Docket No.: 0020-4559P
(PATENT)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Eijiro WATANABE et al.

Application No.: 09/301,766 Confirmation No.: 6045

Filed: April 29, 1999 Art Unit: 1638

For: RAFFINOSE SYNTHASE GENES AND THEIR USE Examiner: D. H. Kruse

APPEAL BRIEF

MS Appeal Brief - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

As required under § 41.37(a), this brief is filed more than two months after the Notice of Appeal filed in this case on December 22, 2006, and is in furtherance of said Notice of Appeal.

The fees required under § 41.20(b)(2) are addressed in the accompanying TRANSMITTAL OF APPEAL BRIEF.

Table of Contents

BRIEF ON BEHALF OF APPELLANT

I. REAL PARTY IN INTEREST

The Assignee of the present application is Sumitomo Chemical Company, Ltd. of Osaka, Japan.

II. RELATED APPEALS, INTERFERENCES, AND JUDICIAL PROCEEDINGS

An Appeal Brief was filed in the copending application no. 08/992,914 on December 26, 2006. An Examiner's Answer has been received and a Reply Brief filed July 16, 2007 in that application.

The copending '914 application is directed to similar subject matter as the present application and the appeal is of the same ground of rejection. That is, both the '914 and the present application present for appeal the question of whether or not a certain degree of sequence identity of a gene or protein sequence is sufficient to establish by the preponderance of the evidence an asserted utility for an invention, and corresponding adequacy of written description and enablement of "how to use" the invention.

III. STATUS OF CLAIMS

The following is the status of the claims as of the mailing of the Final Office Action on August 23, 2006:

Claims 1, 4-10, 16-23, 28 and 29 are pending in the application.

Claims 6 and 7 are indicated as "objected to" in the Office Action Summary and indicated as "allowed" in the Conclusion of the Office Action (paragraph 6 on page 9). Claims 6 and 7 are independent claims and so Appellants suppose that the correct status of claims 6 and 7 is "allowed."

The Examiner's decision rejecting claims 1, 4, 5, 8-10, 16-23, 28 and 29 has been appealed.

Claims 1, 4, 5, 8-10, 16-23, 28 and 29 stand rejected under 35 U.S.C. § 112, first paragraph, for alleged lack of written description support in the specification and also for alleged lack of enablement by the disclosure of the specification.

IV. STATUS OF AMENDMENTS

No further amendments or arguments have been filed pursuant to the Final Office Action of August 23, 2006.

V. SUMMARY OF CLAIMED SUBJECT MATTER

References to page numbering of the specification are made herein to the specification as originally filed.

Claim 1 is directed to an isolated nucleic acid that comprises a polynucleotide encoding an enzyme (raffinose synthase, RFS) that binds a D-galactosyl group through the α(1->6) bond to the hydroxyl group attached to the carbon atom at 6-position of the D-glucose residue in a sucrose molecule to form raffinose. The claimed nucleic acid comprises a polynucleotide comprising a nucleotide sequence selected a group of eight (8) sequences variously described as follows:

(a) a nucleotide sequence encoding the amino acid sequence as depicted in SEQ ID NO: 3,

(b) a nucleotide sequence depicted by the 236th to 2584th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 4,

(c) a nucleotide sequence encoding the amino acid sequence as depicted in SEQ ID NO: 5,

(d) a nucleotide sequence depicted by the 134th to 2467th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 6,

(e) a nucleotide sequence encoding the amino acid sequence as depicted in SEQ ID NO: 7,

(f) a nucleotide sequence depicted by the 1st to 1719th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 8,

(g) a nucleotide sequence obtained from a polynucleotide which is amplified from a nucleic acid obtained from beet with a combination of a PCR primer selected from the group

consisting of SEQ ID NO: 11 and SEQ ID NO: 13 and a PCR primer selected from the group consisting of SEQ ID NO: 12 and SEQ ID NO: 14, wherein said nucleotide sequence hybridizes with a nucleotide sequence complementary to the nucleotide sequence of (a) or (b), in a buffer comprising 0.9M NaCl and 0.09M citric acid at 65°C to 68°C, and

(h) a nucleotide sequence obtained from a polynucleotide which is amplified from a nucleic acid obtained from mustard or rapeseed with a combination of a PCR primer selected from the group consisting of SEQ ID NO: 15, SEQ ID NO: 17 and SEQ ID NO: 19 and a PCR primer selected from the group consisting of SEQ ID NO: 16, SEQ ID NO: 18 and SEQ ID NO: 20, wherein said nucleotide sequence hybridizes with a nucleotide sequence complementary to the nucleotide sequence of any one of (c) to (f), in a buffer comprising 0.9M NaCl and 0.09M citric acid at 65°C to 68°C.

Support for the description of the encoded enzyme as a raffinose synthase having the recited biochemical activity is provided in the specification at, *e.g.* page 2, lines 7-13 and also at page 31, line 22 to page 32, line 4. The latter text also describes by citation to the literature (Lehle et al.) an assay for raffinose synthase activity.

The various polynucleotides (a)-(f) are set forth in the specification at least at page 4, line 19 to page 5, line 13 (numbered as 3. to 9.). The particular sequences are set forth in the Sequence Listing originally filed (and re-filed with a Preliminary Amendment on April 29, 1999).

The polynucleotide (g) is described in Example 4 beginning at page 42, line 10 (as to isolation of a cDNA from beet using PCR). The particular primers of SEQ ID NOS: 11-14 are described in "List 2" at page 13, line 12 and use of these primers to obtain a full length coding sequence of a raffinose synthase cDNA from beet is described at page 53, lines 1-9. The particular Sequence Listing Identifiers correspond to sequences in the Sequence Listing. Hybridization conditions recited in the claim are set forth at page 18, lines 5-10.

The polynucleotide (h) is described at, *e.g.* Examples 6 and 7 beginning at page 45, line 15. The particular primers recited in the claim are described in "List 3" at page 15, line 22 to page 16, line 3, and these primers are described as useful for isolating a cDNA of the complete coding portion of a raffinose synthase gene from mustard or rapeseed at page 53, line 14 to page 54, line 14. The particular Sequence Listing Identifiers correspond to sequences in the Sequence Listing. Hybridization conditions recited in the claim are set forth at page 18, lines 5-10.

Claims 4-10 are directed to particular isolated nucleic acids having specifically recited nucleotide sequences or encoding specifically recited amino acid sequences among the polynucleotides (a) to (f) of claim 1.

Claims 16 - 20 are directed to an isolated nucleic acid comprising a nucleic acid of claim 1, operatively linked to a promoter, vectors comprising a nucleic acid of claim 1 and transformants in which a nucleic acid of claim 1 or such nucleic acid linked to a promoter has been introduced into a host cell. These embodiments are described at, *e.g.* page 6, line 15 to page 7, line 2.

Claims 28 and 29 describe the promoter as one operative in a plant cell or in a yeast cell, respectively. Such promoters are described at, *e.g.* page 29, line 22 to page 30, line 8.

Claims 21 and 22 describe the host cell as a microorganism or plant cell. These embodiments are described at, *e.g.* page 7, lines 3-6 and page 33, line 17.

Claim 23 is directed to a method for producing raffinose synthase by culturing or growing the transformant of claim 18 to produce the raffinose synthase, and collecting the raffinose synthase so produced. Such is described at, *e.g.* page 32, lines 4-20. Purification of raffinose synthase from plants is known in the art, for example as described by Lehle et al. cited at the bottom of page 31.

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

The following grounds of rejection are to be reviewed on appeal:

Claims 1, 4, 5, 8-10, 16-23, 28 and 29 stand rejected under 35 U.S.C. § 112, first paragraph, for alleged lack of written description support in the specification and also for alleged lack of enablement by the disclosure of the specification. (As stated in the Final Office Action at paragraph 4 on page 2 and at paragraph 5 at page 6.)

VII. ARGUMENT

VIIA. Rejections Under 35 U.S.C. § 112, first paragraph – written description

VIIA.1. Claim 1

Claim 1 is rejected under 35 U.S.C. § 112, first paragraph, for alleged lack of written description support of the claimed invention. Appellants respectfully submit that this rejection should be reversed.

In the Final Office Action of August 23, 2006, the Examiner makes a few different assertions regarding the written description requirement. First he states that, as to SEQ ID NO: 7 (note polynucleotides (e) and (f)), description of an incomplete coding sequence does not describe a nucleic acid sequence encoding a raffinose synthase enzyme and that the specification includes no description of the portions of the amino acid sequence necessary for raffinose synthase activity. Second, as to nucleic acids isolated from "beet" or "mustard or rapeseed" (note polynucleotides (g) and (h)), the Examiner asserts that, "merely describing a method by which a nucleic acid may be isolated does not describe the nucleic acid encoding a raffinose synthase as asserted by Applicant." Third, as to SEQ ID NO: 3, and presumably applicable to all of polynucleotides (a), (b) and (e) through (h)[1], the Examiner asserts that a demonstration of the biological activity (and thus of utility) for a protein of a particular amino acid sequence cannot be used to support an assertion of similar activity for a protein of similar sequence.

There are no "bright line" tests for whether or not a specification provides adequate written description of a claimed invention. The Examiner must carefully review the claims, and carefully review the specification to determine whether, in view of what is known in the art at the time the application was filed, the specification provides evidence that the inventor was in "possession" of the invention as claimed. *Capon v. Eshhar,* 76 USPQ2d 1078 (Fed. Cir. 2005); *Faulkner v. Inglis,* 79 USPQ2d 1001 (Fed. Cir. 2006).

[1] The polynucleotides of (c) and (d) are allowed as claims 6 and 7.

As to the Examiner's first assertion, SEQ ID NO: 7 (encoded by nucleotides 1 to 1719 of SEQ ID NO: 8) is indeed only a partial sequence of a raffinose synthase; about 25% of the full-length sequence is missing from the amino-terminal end. However, the instant claim 1 does not recite that the claimed polynucleotide "consists of" the recited sequence. Rather, the claim recites that the polynucleotide "comprises" the recited polynucleotide, and hence also includes any amino acids necessary to complete an amino acid sequence of a raffinose synthase protein. Two such complete amino acid sequences are disclosed in the present application as SEQ ID NOS: 3 and 5. Methods for determining the complete nucleotide sequence of a cDNA encoding raffinose synthase from rapeseed are explained in the specification, as used to obtain complete sequences are obtained for examples from beet and mustard. Alternatively, one of ordinary skill in the art might simply obtain the missing portion of the enzyme from the complete cDNAs for these two proteins that are described (SEQ ID NOS: 2 and 4). The Board is reminded that claim 1 specifically includes as a feature that the encoded protein exhibit a recited enzymatic activity, and so inoperable embodiments are excluded from the claim.

From the above, it is clear that the specification evidences that the inventors had in their possession the invention claimed in claim 1, parts (e) and (f). The Examiner has merely stated a summary conclusion, parroting guidelines to the effect that the specification must set forth an explicit "structure-function relationship"[2] used by the USPTO to implement a policy restricting cloned gene inventions to specifically disclosed species, rather than carefully considering the facts presented by the instant application and claims as required by Federal Circuit case law.

Notwithstanding the failure of the Examiner to carefully consider the facts of the present application, he is simply wrong that the specification does not explain parts of the RFS sequence that should be preserved for activity. The specification explains that certain portions of the amino acid sequence of a RFS should be constrained to high homology to SEQ ID NO: 3 or to SEQ ID NO: 5. See, pp. 20-21, indicating portions of high homology (accounting for both sequences) from amino acids 103 to 213, 255 to 275, 289 to 326 and 609 to 696.

[2] See, pp. 3-4 of the Office Action mailed March 1, 2005, referenced in the Office Action of December 2, 2005, as the "previous Office Action" addressing this issue.

As to the Examiner's second assertion, Federal Circuit case law makes abundantly clear that it is permissible for an Applicant to claim an invention in product-by-process terms. *See, e.g. Enzo Biochem, Inc. v. Gen-Probe, Inc.,* 63 USPQ2d 1609 (Fed. Cir. 2002) and *Fiers v. Revel,* 25 USPQ2d 1601, 1605 (Fed. Cir. 1993). The Examiner's position with respect to parts (g) and (h) of claim 1 is simply complete legal error.

The Examiner's third argument is in the first place more related to the issue of enablement of the utility of the invention than to the written description of the structure of the invention. More substantively, the Examiner's third argument is contrary to the substantial evidence in the record of the present application. The present specification describes in detail a method for cloning raffinose synthase (RFS) genes from plants of broadly diverse genera (*Glycine, Beta, Brassica*). The specification describes using sequence information of a cDNA encoding part of a RFS from *Glycine max* (SEQ ID NO: 2) to prepare a set of PCR primers that will hybridize to a degenerate set of sequences[3] that are used to amplify mRNA obtained from other plants and so isolate fragments of RFS cDNA. These initial amplification products are sequenced, then that further data are used to prepare a new set of primers specific for RFS for the particular plant being studied. The second set of primers is used to make new amplification products that are cloned and from which the complete sequence of the cDNA is obtained (*see,* the Examples 1-6 of the specification). This approach was used three times in working examples of the present specification to successfully obtain RFS cDNAs from three different plants. The complete coding portions of the cDNA for *Beta vulgaris* and *Brassica juncea* (SEQ ID NOS: 4 and 6, respectively) and part of the coding portion of a cDNA from *Brassica napus* (SEQ ID NO: 8) are presented. Appellants have presented evidence in the form of a Declaration of Dr. Watanabe that demonstrates unequivocally that a protein having the amino acid sequence of SEQ ID NO: 5 has biological activity of a RFS, and this is not disputed by the Examiner[4]. Therefore, plainly the approach described in the specification can be used successfully to isolate a cDNA encoding RFS from diverse genera of plants.

[3] The primers used for initial amplification include degenerate positions or inosine bases that recognize A/G or C/T alternatives.

[4] Claims 6 and 7, directed to this embodiment (also represented by (c) and (d) in claim 1) are allowed.

The Examiner asserts that one of ordinary skill in the art is not able to distinguish RFS enzymes from the closely related stachyose synthase enzymes, and therefore the actual biochemical activity of the proteins encoded by the cDNAs of SEQ ID NOS: 4 and 8 remain unknown. Accordingly, the Appellants are asserted to have provided only one example of a RFS-encoding cDNA and therefore, except for the claims directed to the species having proven activity, the written description of the invention is inadequate.

RFS enzymes are a subfamily of enzymes grouped together with the subfamily of stachyose synthases (STS) in a family of glycoside hydrolase enzymes. Appellants have presented substantial evidence that one of ordinary skill in the art can distinguish RFS from STS members of the glycoside hydrolase family. This evidence is in the form of the data in Tables 1 and 2 and Figure 1 presented with Appellants' Amendment filed February 11, 2004 and in Table 3 and Exhibit 1 presented with Appellants' Amendment filed November 15, 2004. These data, and the Exhibit supporting the robustness of the analysis, show that RFS enzymes are distinguishable from STS enzymes by determination of the degree of sequence identity to SEQ ID NO: 1, 3, 5 or 7 according to the present specification. In particular, the data show that RFS enzymes among themselves are at least 50% identical at the amino acid level, and that STS enzymes are similarly homologous to each other (actually a bit more so, about 65%). However, the degree of identity between RFS and STS enzymes is at most about 45%. Sequence identity analysis permits the artisan of ordinary skill to illustrate the distinction between RFSs and STSs by a "dendrogram", as shown in Figure 1 attached to the Amendment filed February 11, 2004.

Furthermore, Appellants have argued that the specification of the copending application 08/992,914, which discloses additional examples of RFS cDNAs cloned essentially in the manner described in the present specification, provides another example in which biological activity of a RFS cDNA[5] is demonstrated, and that this demonstration further evidences the effectiveness of the methods described in the present specification in obtaining cDNAs encoding RFS proteins. (*See,* Appellants amendment of June 2, 2006, at page 8, lines 3 ff.) The Examiner

[5] from *Vinca faba*, SEQ ID NO: 1 of the '914 application.

has so far dismissed this further evidence because, "each application is to be considered on its own merits." (*See,* page 5, lines 17-18 of the Final Office Action.) That may be the case, but the facts of the existence of the '914 application and its working examples may be considered as evidence in this application.

In order to meet the requirements for adequate written description, the specification must provide evidence that the inventors "possessed" the invention as claimed at the time the application was filed. *Vas-Cath v. Murhurkar* 19 USPQ2d 1111 (Fed. Cir. 1991). The evidentiary standard that must be met by Appellants is only that of the preponderance of the evidence. *See, In re Oetiker*, 24 USPQ2d 1443, 1444 (Fed. Cir. 1992). The Examiner seems to be improperly requiring that Appellants meet a higher evidentiary burden, *i.e.* "clear and convincing" evidence or even "beyond reasonable doubt."

Appellants submit that the evidence of record in the present application firmly establishes that it is "more likely than not" that all of the sequences disclosed in the present application are those of RFS enzymes. This has been unequivocally established by biochemical assay for one disclosed sequence, and sequence similarity as analyzed by one of ordinary skill in the art establishes that it is more likely than not true for the others. Furthermore, the same approach for cloning RFS-encoding cDNAs used in the present application has been further applied by the Appellants, as described in a copending application, and yet a further demonstration that the approach obtains cDNA encoding an RFS enzyme, as determined by assay of another expressed cDNA, has been made.

Appellants submit that the present specification, by showing reduction to practice of four species of the claimed invention obtained from three diverse genera of plants, adequately evidences that the inventors "had possession" of the claimed invention at the time the present application was filed. Accordingly, the decision of the Examiner that claim 1 is not supported by adequate written description in the specification should be **reversed**.

VIIA.2. Claims 4 and 5

Claim 4 of the present application is directed to an isolated nucleic acid comprising a nucleotide sequence encoding the amino acid sequence of SEQ ID NO: 3. Claim 5 recites a specific polynucleotide sequence of SEQ ID NO: 4 from which SEQ ID NO: 3 is derived. SEQ ID NO: 3 is the amino acid sequence of the complete coding region of a cDNA obtained from a Chenopdiaceae plant (*Beta vulgaris*) in Examples 3 and 4. As the sequence is of a complete protein coding portion of the cDNA, the rejection of these claims is based only upon the argument of the Examiner regarding the evidence that the encoded enzyme has RFS activity.

Appellants have explained above that there is substantial evidence in the record that establishes at least to the preponderance of the evidence standard that the amino acid sequence of SEQ ID NO: 3, encoded by the cDNA of SEQ ID NO: 4, represents a protein having RFS activity. The particular amino acid sequences in question were determined by a cloning method generally accepted in the art as useful for cloning functionally homologous proteins across species lines.

Finally, claims 4 and 5 recite specific sequences at either the nucleotide or amino acid level. In either case, the skilled artisan can readily determine the exact structure or family of structures encompassed by the claims and so there is no question that the inventors "possessed" the inventions described in these two claims.

For all of the above reasons, the rejection of claims 4 and 5 under 35 U.S.C. § 112, first paragraph, for alleged lack of adequate written description in the specification, should be **reversed**.

VIIA.3. Claims 8 and 9

Claim 8 of the present application is directed to an isolated nucleic acid comprising a nucleotide sequence encoding the amino acid sequence of SEQ ID NO: 7. Claim 9 recites a specific polynucleotide sequence of SEQ ID NO: 8 from which SEQ ID NO: 7 is derived. SEQ

ID NO: 7 is the amino acid sequence of part of the coding region of a cDNA obtained from a Cruciferae plant (*Brassica napus*) in Examples 6 and 7.

Among the arguments presented by the Examiner against claim 1, only the issues of a partial sequence and the sufficiency of the evidence that a RFS enzyme is encoded are applicable to claims 8 and 9.

Appellants have explained above that claims 8 and 9 are directed to nucleic acids comprising the recited sequences, and thus may include additional nucleotides encoding further amino acids as may be necessary to provide an enzyme having RFS activity. Appellants have explained previously that the specification provides description of two complete RFS enzyme amino acid sequences, and that these data can be used to prepare the additional sequences for attachment to the partial sequence of SEQ ID NO: 7. The specification also describes regions of high homology that should be present in an enzyme having RFS activity. For the convenience of the Board, Appellants present as Exhibit 4 attached hereto an alignment of the amino acid sequences of SEQ ID NO: 7 (sc-07) at issue with SEQ ID NO: 5 (sc-05), which represents a protein demonstrated by biochemical assay to have activity as a raffinose synthase.[6] In the alignment, identical amino acids are shown by *. The regions of high homology within raffinose synthases described in the specification are indicated by the shaded portions of the sequence. Appellants submit that the missing 4% of that region (or for that matter the entirety of the missing amino-terminal portion) may be supplied by the corresponding amino-terminal end sequences of SEQ ID NO: 3 or 5 as desired by the practitioner of the invention.

Appellants have also explained above that the evidence of record in the present application is sufficient, at least to the standard of the preponderance of the evidence, to establish that the amino acid sequence of SEQ ID NO: 7 is that of a RFS enzyme.

[6] Appellants submit that Exhibit 4 does not constitute "new evidence" as it merely presents data actually present in the record in the form of SEQ ID NOS: 5 and 7 and otherwise described in the specification. However, Exhibit 4 presents this information in a different format convenient for consideration by the Board.

Finally, claims 8 and 9 recite specific sequences at either the nucleotide or amino acid level. In either case, the skilled artisan can readily determine the exact structure or family of structures encompassed by the claims and so there is no question that the inventors "possessed" the inventions described in these two claims.

For the reasons above, the decision of the Examiner rejecting claims 8 and 9 under 35 U.S.C. § 112, first paragraph, for lack of adequate written description support by the specification, should be **reversed**.

VIIA.4. Claim 10

Claim 10 is directed to an isolated nucleic acid comprising the entirety of SEQ ID NO: 4, SEQ ID NO: 6 or SEQ ID NO: 8. The scope of claim 10 differs slightly from that of claims 5, 7 and 9 in that the entire length of the recited nucleotide sequence is recited in claim 10. In contrast, claims 5, 7 and 9 recite the coding portions of the nucleotide sequences.

Appellants's arguments above apply to claim 10 as well.

Claim 10 recites a group of specific structures that are expressly stated in the Sequence Listing as originally filed. There can be no doubt that the specification describes these sequences exactly and so no doubt that claim 10 meets the requirement for written description of the claimed invention.

For the reasons above, the decision of the Examiner rejecting claim 10 under 35 U.S.C. § 112, first paragraph, for lack of adequate written description support by the specification, should be **reversed**.

VIIA.5. Claims 16-23, 28 and 29

Claims 16-23, 28 and 29 are dependent ultimately from claim 1 and stand rejected for the same reasons as claim 1 is rejected. These dependent claims are directed to embodiments of the invention in which a nucleic acid providing a sequence encoding a RFS enzyme is operatively

linked to a promoter (claim 16) or placed into a vector (claim 17) or to a transformed host cell comprising the nucleic acid of claim 1 either *per se,* or as part of a promoter-structural gene construct or as part of a vector (claims 18, 19 and 20, respectively). Claims 22, 23, 28 and 29 further define the nature of the host cell or the nature of the promoter, respectively.

The Examiner has so far presented no reason for rejection of claims 16-23, 28 and 29 independent from the rejection of claim 1. Thus, the Board is respectfully requested to consider that, should the decision of the Examiner with respect to any part (a) through (h) of claim 1 be reversed, the dependent claims 16-23, 28 and 29 should be indicated as allowable if rewritten to recite the allowable part of claim 1.

VIIB. Rejections under 35 U.S.C. § 112, first paragraph – enablement

Claims 1, 4, 5, 8-10, 16-23, 28 and 29 stands rejected under 35 U.S.C. § 112, first paragraph, for alleged lack of enablement of the claimed invention by the disclosure of the specification. The Examiner's position on this issue is essentially that grouping of enzyme primary structures into families based upon sequence identity is insufficient to support an assertion that a protein of unconfirmed activity will have the activity ascribed to that family demonstrated by biochemical assay of at least one of its members. The Examiner therefore asserts that the present specification is enabling of "how to use the invention" only for those proteins for which activity as a raffinose synthase is actually demonstrated by biochemical assay. In the present instance, the Examiner asserts that the claims must be limited with respect to the amino acid sequence of the enzyme to SEQ ID NO: 6, which is the sole amino acid sequence described in the present specification for which raffinose synthase activity has been actually demonstrated.

Appellants disagree.

VIIB.1. Claim 1

The question of enablement is to be considered under a multifactor analysis as set forth in *In re Wands*, 8 USPQ2d 1400 (Fed. Cir. 1988). It is incumbent upon the Examiner to first establish a *prima facie* case for lack of enablement. *In re Wright*, 27 USPQ2d 1510, 1513 (Fed. Cir. 1993) (holding examiner must provide a reasonable explanation as to why the scope of protection provided by a claim is not adequately enabled by the disclosure). Should the Examiner do so, then the Appellant must establish, by the preponderance of the evidence, that no undue experimentation is required to practice the invention as claimed. *In re Oetiker*, 24 USPQ2d at 1444.

Appellants first submit that the Examiner has never established a proper *prima facie* case for lack of enablement of the claimed invention. Proper analysis of the question of enablement requires that the factors of 1) the breadth of the claims, 2) the nature of the invention, 3) the level of ordinary skill in the art, 4) the amount of experimentation needed, 5) the state of the art at the time the invention was made, 6) the amount and quality of guidance provided by the specification, 7) the presence or absence of working examples and 8) the predictability in the art. Of these factors, the Examiner repeatedly has only addressed the predictability in the art. The Examiner's position is that, because there is evidence in the record for RFS activity only for a protein of amino acid sequence of SEQ ID NO: 5, and the degree of sequence identity among the amino acid sequences identified in the working examples is as low as 50%, Appellants cannot reliably assign the biochemical activity of a raffinose synthase to the amino acid sequences of SEQ ID Nos: 3 and 7.

Applicants note first that analysis of enablement is a question of whether "undue experimentation" is required to practice the invention throughout its claim scope. Consideration of the question of undue experimentation is by weighing all of several factors enumerated in *In re Wands,* 8 USPQ2d 1400 (Fed. Cir. 1988).

The Examiner fails to meet his burden of establishing a *prima facie* lack of enablement. The Examiner's analysis of the question of undue experimentation looks only at the factor of

whether working examples of the claimed invention are described in the specification and an assertion that it is unpredictable whether any particular nucleic acid produced according to the teachings of the invention would in fact exhibit raffinose synthase activity. This analysis is legally insufficient to establish *prima facie* lack of enablement, as the Examiner fails to consider the breadth of the claims, the nature of the invention, the level of ordinary skill in the art, the quantity of the experimentation needed, the guidance provided by the specification (other than the presence or absence of working examples) and the state of the art at the time the invention was made. Furthermore, the kind of predictability, *a priori* knowledge of functionality of the enzyme obtained using the methods of the invention, is not the kind of predictability envisioned by the Court in *Wands*. The instant rejection cannot properly be sustained against claim 1.

The nature of the invention and the breadth of the claims

The claimed invention relates to isolated nucleic acids that encode an enzyme having a defined biological activity. As to claim 1, the invention as most broadly stated (*e.g.* (g) or (h)) lies in a nucleic acid that is defined by (1) inclusion of at least certain sequence features (that is, the PCR primer sequences that are used to generate the claimed nucleic acid), (2) hybridization to a certain reference sequence and (3) encoding a protein having a defined enzymatic activity. Claim 1 includes narrower definitions (a) through (f), related to particular amino acid sequences. Among descriptions (a) through (f), the nucleic acid as described by amino acid sequence SEQ ID NO: 5 ((c) and (d)) is proven to encode a protein having RFS activity, the nucleic acid described by reference to SEQ ID NO: 3 ((a) and (b)) represents the entire coding sequence of a RFS protein and the nucleic acid described by reference to SEQ ID NO: 7 ((e) and (f)) represents about 70% of the length of the coding sequence of a RFS protein.

Inoperative embodiments are excluded from the claims by the requirement that the encoded protein have RFS activity.

The art of molecular biology, in particular the art of expression of recombinant proteins, is one in which the artisan of ordinary skill expects to perform a few weeks or months of experimentation in generating variants of a protein, then isolating clones encoding those

variants and then (perhaps) re-cloning the isolated variants into vectors for expressing a protein, and then screening expressed proteins for activity.

The level of ordinary skill in the art

The artisan of ordinary skill in the art of cloning and expressing recombinant proteins is generally accepted as one having a Ph.D. degree and perhaps higher *i.e.*, having significant post-doctoral laboratory experience. Such a person is skilled in the design and performing of experiments for isolating DNA clones and for screening them for a desired property, for example encoding a protein having a particular activity.

The amount of experimentation needed

The amount of experimentation needed to practice the present invention is not unduly large or burdensome. The practitioner must isolate a template genomic DNA or RNA from an organism, perform a polymerase chain reaction using primers described in the specification to generate an amplified fragment, clone that fragment into an expression vector, express the encoded protein and then screen the protein for activity as a raffinose synthase. All of these steps are either well-known in the art or described in detail in the specification (*e.g.* pp. 31-33 (bacterial expression of the cloned cDNA and assay for RFS activity and Examples 1-6 beginning at p. 38) and furthermore are expected to be performed by the artisan of ordinary skill.

The state of the art at the time the invention was made

At the time the invention was made, the state of the art of molecular biology was such that the various laboratory operations that must be performed to carry out the experimentation required to practice the instant invention, *i.e.* cloning of DNA molecules and expressing them in a host cell, were routine. Also, polymerase chain reaction amplification of nucleic acids was routine.

The raffinose content of a number of organisms, especially including plants and some algae, was known. The biochemistry of raffinose synthesis in plants had been established, and the role of raffinose synthases as rate-limiting of raffinose production was known. (*See, e.g.* pp. 1-2 of the specification.)

A biochemical assay for raffinose synthase activity was described. See Lehle et al., *Eur. J. Biochem.* 38:103 (1973) (attached).

The guidance provided by the specification including the presence or absence of working examples

The specification provides ample guidance to the skilled artisan for practicing the invention broadly. In particular, the specification discloses in detail how to clone DNAs encoding putative raffinose synthase enzymes. The specification provides details such as organisms likely to be useful for isolating template genomic DNA or RNA from plants commensurate in scope with claim 1 and corresponding PCR primers (Chenopdiceae (for beet), p. 11, line 14; Cruciferae (for mustard and rapeseed), p. 13, line 18 and associated PCR primers in Lists 2 and 3).[7] The specification describes methods for cloning DNA encoding a putative raffinose synthase enzyme from an RNA fraction, including an extensive list of primers that can be utilized for PCR amplification from templates obtained from different organisms (*see, e.g.* Lists 6 and 7 at p. 43; Lists 8 and 9 at page 46; List 10 at p. 47). The specification describes methods for expressing the cloned DNA in plant cells and in bacteria (*see, e.g.* pages 29 to 37). The specification describes a biochemical assay for raffinose synthase, referring to the Lehle article noted above and summarizing the procedure beginning at the bottom of page 31.

The specification also provides a number of working examples of isolation of partial or complete raffinose synthase genes from a number of different plants (*see*, Examples 1-7) and of creation of an expression vector for use in plants (Example 8) transformation of a plant (mustard) with a cloned DNA encoding a raffinose synthase (Example 9).

[7] The specification also includes information useful for obtaining RFS cDNA from soybean.

The predictability in the art

The Examiner asserts that the art of recombinant DNA cloning and recombinant protein expression is unpredictable. The Examiner argues that a practitioner of the invention must engage in trial and error experimentation to identify cloned DNAs that encode functional raffinose synthase genes.

The Examiner's argument is simply incorrect. First, the skilled artisan can follow detailed teachings in the specification of how to clone, express and evaluate DNAs that are likely to encode functional raffinose synthase enzymes. It is true that it is a bit unpredictable whether any individual clone made in an experiment will include a DNA encoding a functional enzyme, but it is not unpredictable whether the skilled artisan would succeed in identifying at least one functional DNA in an experiment as a whole. To the contrary, it is very likely that the skilled artisan would find a cloned DNA encoding a functional enzyme by following the teachings of the specification. Appellants note that the experimental approach described in the specification resulted in identification of four cDNAs described in this application and additional cDNAs as described in the copending '914 application.

The Board might consider certain details from the *Wands* case. In *Wands*, an invention related to isolation of hybridomas that secreted a particular antibody was deemed broadly enabled despite that extensive screening of many cloned cell lines was necessary AND that the success rate of the screening was only 2.8%, including experiments that failed to generate any operable clones at all. The *Wands* panel expressly stated that experimentation, such as the cloning and screening experiments described in the present application, that is expected to be performed by the artisan of ordinary skill, is not undue experimentation.

Applicants submit that a proper weighing of the *Wands* factors will lead the Board to a proper conclusion that no undue experimentation is required to practice the present invention as claimed in claim 1. Accordingly, the Examiner's decision rejecting claim 1 for lack of enablement should be **reversed** because the Examiner failed to establish a *prima facie* lack of enablement.

Furthermore, Appellants have provided evidence, in the form of the Watanabe Declaration attached to their Amendment of February 11, 2004, to support an assertion that the procedures described in the specification result in cloning of cDNAs encoding RFS enzymes. Appellants have also provided evidence that one of ordinary skill in the art can readily distinguish a RFS from a STS or another class of closely related proteins, Seed Imbibation Proteins (SIPs). The data in Figure 1 attached to Appellants' Amendment of February 11, 2004, and submitted as part of the Nagasawa Declaration (copied from the '914 appplication file and submitted with Appellants' Amendment of June 2, 2006) demonstrates unequivocally that the RFS subfamily of glycoside hydrolases (see Appellants' discussion of Peterbauer et al., below) is easily distinguished from the STS or SIP subfamilies of glycoside hydrolases on the basis that RFSs are more similar to each other, and STSs are more similar to each other, than RFSs are similar to STSs. This relationship among their amino acid sequences can be used to construct a "molecular phylogenetic tree" upon a branch of which any particular amino acid sequence thought to represent a RFS or STS (or SIP) can be placed. The Nagasawa Declaration further explains that this analysis is robust in its conclusions (though perhaps the specific degrees of sequence similarity may vary) to three different approaches to sequence similarity analysis.

The Examiner has attempted to support his position regarding unpredictability in the art with evidence from the scientific literature. The Examiner has cited Richmond et al. *Plant Physiology* (2000) and Duggleby, *Gene* (1997) for a general assertion that, "The art teaches that one of skill in the art cannot assume the function of the polypeptide encoded by an isolated nucleic acid solely based on sequence similarity to a known polypeptide sequence. The Examiner cites Peterbauer et al., *Planta* (2002) for the proposition that RFS enzymes have high sequence homology to STSs and SIPs. The Examiner cites Peterbauer et al., *Planta* (1999) for the proposition that their group was the first to isolate a nucleic acid encoding a STS protein. See, the Office Action of August 11, 2003 at p. 3. The Examiner cites Bowie et al., *Science* (1990) for the proposition that it is the three dimensional structure of an enzyme that confers its activity and that folding of a protein can be sensitive to minor changes in sequence. The Examiner cites Lazar *Mol. Cell. Biol.* (1998) as an example of an instance in which a certain change in sequence to TGF-α had no effect on the protein, but another instance of amino acid

substitution "sharply reduced biological activity". The Examiner cites Broun et al., *Science* (1998) for the proposition that a few amino acid substitutions can have radical effects on the activity of an enzyme. *See,* the Office Action of November 20, 2002, at pp. 6-7.

Appellants do not dispute the general conclusion reached from the Bowie, Lazar and Broun papers that it is the three-dimensional structure of a protein that confers its biological activity, or that sometimes there are particular amino acids that must be conserved in the linear sequence to preserve the correct folding of the protein, or even that in some instances two distinct enzymes will share extensive portions of amino acid sequences. These concepts are well-known to the molecular biologist of ordinary skill in the art and they do suggest that it is somewhat unpredictable whether mutating a protein will result in maintaining, lessening or improving its biological activity. However, this is not determinative of whether undue experimentation is required to practice the instant invention. All that such unpredictability establishes is that, without actual assay data, one cannot say beyond reasonable doubt that a mutated protein will retain its original activity. However, this is not the proper standard of evidence to consider during patent prosecution. Appellants' burden is to only establish that it is more likely than not that the proteins of amino acid sequences 3 and 7 represent a protein having RFS activity, or that a cDNA obtained as described in parts (g) and (h) of claim 1 encodes such a protein.

The Examiner asserts that Richmond et al. indicates that more than sequence similarity is needed as evidence of function, pointing out the paragraph bridging the left and right columns of page 497. Appellants see here only a description of domains present in members of the cellulose synthase family of proteins. Indeed, Richmond might be interpreted as more supportive of Appellants' position that sequence similarity is a useful tool for grouping proteins by activity. The Board might take note of Figure 1 of the paper, showing assignment of members of the family to subfamilies CesA, CesB, CesD, etc. based upon a molecular phylogeny. The Board may usefully compare Figure 1 of Richmond with Figure 1 attached to the Nagasawa Declaration, which shows a similar molecular phylogeny among RFSs, STSs and a SIP, with the result of clear separation of the three groups of enzymes.

The Examiner points out the last paragraph of Duggleby. There, the author states, "Ultimately the function of any DNA sequence, whose identity is based solely on homology, can only be proven by experiments designed to evaluate that function." Again, this simply goes to the standard of the proof. For purposes of alleging utility in a patent application, the standard of proof is merely the preponderance of the evidence. Appellants note that Duggleby has no problem asserting function from sequence similarity. The Board might consider the text of the Note Added In Proof: "Recent examination of GenBank expressed sequence tags has identified three sequences … that may represent higher plant ALS small subunits. The last of these gives a very good match to the *P. purpurea* sequence; over residues 83-154 there are 46 identical, and 10 similar, amino acids." The Board might further note that the author's conclusion is based upon a degree of identity of only 71% at the amino acid level of a partial amino acid sequence.

The Duggleby paper describes study of the small subunit of the acetolactate synthase (ALS) from a bacterium, yeast and an alga. The paper provides an alignment of the genes from these three organisms (Figure 2). The authors note that there is only "limited similarity" among the three sequences, but nonetheless were able to detect a number of known bacterial ALS genes and also discovered the eukaryotic versions of the gene using a BLAST search of GENBANK and the bacterial sequence (*B. flavum*) as a query. See, p. 247, under Results and Discussion. Thus, Duggleby in fact also supports Appellants' assertion that comparison of sequence data is a common technique in the art for predicting biochemical function of a protein. ("These results clearly indicate that *S. cerevisiae* and *P. purpurea* contain a gene that could encode an ALS small subunit." (at the top of the right column on p. 247.))

Peterbauer (2002) describes isolation of a raffinose synthase gene from *P. sativum* (pea). The Examiner asserts that Peterbauer teaches that RFSs, STSs and SIPs demonstrate high overall sequence homology. This has not been disputed by Appellants. Peterbauer discusses this result in terms of assignment of all three of these enzyme types to the glycoside hydrolase enzyme family (p. 841, right column, above Figure 1). Appellants' argument is that RFSs are more alike, and STSs are more alike, than RFSs resemble STSs and therefore these members of the glycoside hydrolase family are distinguishable subfamilies.

Appellants note that the Examiner has read Peterbauer (2002) rather selectively. At the top of the right column on p. 841, Peterbauer easily distinguishes a STS transcript from a RFS transcript on the basis of sequence identity.

In fact, Peterbauer uses an approach to cloning the pea RFS gene that is similar to that described in the present specification. That is, PCR primers designed from the amino acid sequence of the RFS were used to amplify template DNA from the pea plant. Then the resulting cDNA was expressed in a cell and the protein so produced was assayed for RFS activity. These teachings may usefully be compared with the working examples 1-6 of the present specification and the Watanabe Declaration.

Peterbauer (2002) does not particularly support the Examiner's position. The authors note that, "to distinguish between raffinose synthase and stachyose synthase, the primers were chosen to encompass a block of about 80 amino acids, which is exclusively present in stachyoses synthases." (Top of page 841, right column.) This establishes that there are in fact amino acid sequence elements that serve to distinguish a RFS from a STS. Second, the Examiner has read the paper very selectively, urging the data showing sequence similarity, but ignoring for example, the text at the top of the right column of p. 841, "To isolate a cDNA encoding for raffinose synthase by reverse transcription-PCR, degenerate primers were designed based upon amino acid motifs conserved among *Cucumis sativa* raffinose synthase, stachyose synthase and related sequences." Thus, Peterbauer et al. were satisfied that they could reliably distinguish among such sequences either by biochemical or sequence analysis methods.

Thus, none of the papers proffered by the Examiner in rebuttal of Appellants' arguments is effective to undermine either their argument that the specification is enabling of practice of the invention, or the evidence of the Nagasawa Declaration that one of ordinary skill in the art can readily determine by amino acid sequence analysis whether a given amino acid sequence represents a RFS, a STS or a SIP.

Since the Examiner has in the first instance failed to establish a *prima facie* lack of enablement of the claimed invention, and in the second instance has failed to effectively rebut

Appellants' arguments and evidence offered in support of enablement of the claimed invention, the present rejection of claim 1 under 35 U.S.C. § 112, first paragraph, for alleged lack of enablement, must be **reversed**.

VIIB.2. – claim 4

Claim 4 is directed to isolated nucleic acids encoding the amino acid sequence of SEQ ID NO: 3. All of Appellants' arguments against the Examiner's rejection of claim 1 for lack of enablement are applicable as well to claim 4. However, the breadth of this claim is substantially narrower than the breadth of claim 1. Also, the amino acid sequence of SEQ ID NO: 3 is of the complete length of the protein and the degree of sequence identity to SEQ ID NO: 5, proven to represent an enzyme having RFS activity in the Watanabe Declaration, is 63%, substantially higher than the degree of identity between a RFS and STS (*see*, Table 2 attached to Appellants' Amendment of February 11, 2004). Therefore, the degree of unpredictability as to whether SEQ ID NO: 3 encodes a RFS enzyme or not may be considered to be lower than that for claim 1 as a whole, and so enablement of claim 4 should be weighed separately from enablement of claim 1.

For all of the reasons above, Appellants urge that the Examiner's decision that the specification fails to enable claim 4 should be **reversed**.

VIIB.3. – claim 5

Claim 5 is directed to an isolated nucleic acid comprising a recited portion of sequence of SEQ ID NO: 4. All of Appellants' arguments against the Examiner's rejection of claim 1 for lack of enablement are applicable as well to claim 5. However, the breadth of this claim is substantially narrower than the breadth of claim 1. Also, the recited portion of SEQ ID NO: 4 encodes the complete length of SEQ ID NO: 3, a full-length RFS having a degree of sequence identity to SEQ ID NO: 5, proven to represent an enzyme having RFS activity in the Watanabe Declaration, of 63%, substantially higher than the degree of identity between a RFS and STS (*see*, Table 2 attached to Appellants' Amendment of February 11, 2004). Therefore, the degree

of unpredictability as to whether SEQ ID NO: 4 encodes a RFS enzyme or not may be considered to be lower than that for claim 1 as a whole.

Furthermore, the only experimentation necessary to determine conclusively whether the sequence SEQ ID NO: 4 in fact does encode a RFS enzyme is to clone this sequence into an expression vector, transform a bacterial or plant host cell with the vector and test the transformed bacteria or plant tissue for expression of RFS activity in the manner described in the specification. (*See*, *e.g.* pp. 31-37 of the specification.) Such experimentation must be considered well-guided by the specification and expected by the artisan of ordinary skill, and so not "undue". Accordingly, enablement of claim 5 should be weighed separately from enablement of claim 1.

For all of the reasons above, Appellants urge that the Examiner's decision that the specification fails to enable claim 5 should be **reversed**.

Also, the specification, at page 26, line 13 to page 28, line 21, describes use of nucleic acids of the invention in genotyping analysis or for detection of mutation in raffinose synthase genes or for marking cloned plant varieties. These utilities are independent of whether or not the cloned DNA encodes a protein having raffinose synthase activity, for example, a nucleic acid encoding only a part of a raffinose synthase gene is adequate for use in such methods. At least for genotyping and plant variety identification even nucleic acids unrelated to raffinose synthase genes are useful. Therefore, the Board should consider that the specification provides adequate description of how to use the nucleic acid of claim 5 and the Examiner's decision to the contrary may by **reversed** for this reason alone.

VII.B.5 – Claim 8

Claim 8 is directed to isolated nucleic acids encoding the amino acid sequence of SEQ ID NO: 7. All of Appellants' arguments against the Examiner's rejection of claim 1 for lack of enablement are applicable as well to claim 8. However, the breadth of claim 8 is substantially narrower than the breadth of claim 1. Therefore, the degree of unpredictability as to whether

SEQ ID NO: 8 encodes a RFS enzyme or not may be considered to be lower than that for claim 1 as a whole, and so enablement of claim 8 should be weighed separately from enablement of claim 1.

VII.B.6 – Claim 9

Claim 9 is directed to an isolated nucleic acid comprising a recited portion of sequence of SEQ ID NO: 8. All of Appellants' arguments against the Examiner's rejection of claim 1 for lack of enablement are applicable as well to claim 9. However, the breadth of this claim is substantially narrower than the breadth of claim 1. Therefore, the degree of unpredictability as to whether SEQ ID NO: 8 encodes a RFS enzyme or not may be considered to be lower than that for claim 1 as a whole, and so enablement of claim 9 should be weighed separately from enablement of claim 1.

Furthermore, the only experimentation necessary to determine conclusively whether the sequence SEQ ID NO: 8 in fact does encode a RFS enzyme is to clone this sequence into an expression vector, transform a bacterial or plant host cell with the vector and test the transformed bacteria or plant tissue for expression of RFS activity in the manner described in the specification. (*See*, *e.g.* pp. 31-37 of the specification.) Such experimentation must be considered well-guided by the specification and expected by the artisan of ordinary skill and so not "undue". Accordingly, the Board should weigh enablement of claim 9 separately from enablement of claim 1.

For all of the reasons above, Appellants urge that the Examiner's decision that the specification fails to enable claim 9 should be **reversed**.

Also, the specification, at page 26, line 13 to page 28, line 21, describes use of nucleic acids of the invention in genotyping analysis or for detection of mutation in raffinose synthase genes or for marking cloned plant varieties. These utilities are independent of whether or not the cloned DNA encodes a protein having raffinose synthase activity, for example, a nucleic acid encoding only a part of a raffinose synthase gene is adequate for use in such methods. At least

for genotyping and plant variety identification even nucleic acids unrelated to raffinose synthase genes are useful. Therefore, the Board should consider that the specification provides adequate description of how to use the nucleic acid of claim 9 and the Examiner's decision to the contrary may by **reversed** for this reason alone.

VII.B.7 – Claim 10

Claim 10 is directed to an isolated nucleic acid comprising the entirety of any one of SEQ ID Nos: 4, 6 or 8. All of Appellants' arguments against the Examiner's rejection of claim 1 for lack of enablement are applicable as well to claim 10. However, the breadth of this claim is substantially narrower than the breadth of claim 1. Therefore, the degree of unpredictability as to whether SEQ ID Nos: 4 and 8 encode a RFS enzyme or not may be considered to be lower than that for claim 1 as a whole, and so enablement of claim 10 should be weighed separately from enablement of claim 1.

Furthermore, the only experimentation necessary to determine conclusively whether the sequences SEQ ID NO: 4 and 8 in fact do encode a RFS enzyme is to clone these sequences into an expression vector, transform a bacterial or plant host cell with the vector and test the transformed bacteria or plant tissue for expression of RFS activity in the manner described in the specification. (*See*, *e.g.* pp. 31-37 of the specification.) Such experimentation must be considered well-guided by the specification and expected by one of ordinary skill in the art and so not "undue". Accordingly, the Board should weigh enablement of claim 10 separately from claim 1.

For all of the reasons above, Appellants urge that the Examiner's decision that the specification fails to enable claim 10 should be **reversed**.

Also, the specification, at page 26, line 13 to page 28, line 21, describes use of nucleic acids of the invention in genotyping analysis or for detection of mutation in raffinose synthase genes or for marking cloned plant varieties. These utilities are independent of whether or not the cloned DNA encodes a protein having raffinose synthase activity, for example, a nucleic acid

encoding only a part of a raffinose synthase gene is adequate for use in such methods. At least for genotyping and plant variety identification even nucleic acids unrelated to raffinose synthase genes are useful. Therefore, the Board should consider that the specification provides adequate description of how to use the nucleic acid of claim 10 and the Examiner's decision to the contrary may by **reversed** for this reason alone.

VII.C.8 – Claims 16-23, 28 and 29

Claims 16-23, 28 and 29 are dependent ultimately from claim 1 and stand rejected for the same reasons as claim 1 is rejected. These dependent claims are directed to embodiments of the invention in which a nucleic acid providing a sequence encoding a RFS enzyme is operatively linked to a promoter (claim 16) or placed into a vector (claim 17) or to a transformed host cell comprising the nucleic acid of claim 1 either *per se,* or as part of a promoter-structural gene construct or as part of a vector (claims 18, 19 and 20, respectively). Claims 22, 23, 28 and 29 further define the nature of the host cell or the nature of the promoter, respectively.

The Examiner has so far presented no reason for rejection of claims 16-23, 28 and 29 independent from the rejection of claim 1. Thus, the Board is respectfully requested to consider that, should the decision of the Examiner with respect to any part (a) through (h) of claim 1 be reversed, the dependent claims 16-23, 28 and 29 should be indicated as allowable if rewritten to recite the allowable part of claim 1.

VII.C. – Summary and Conclusion

Claims 1, 4, 5, 8-10, 16-23, 28 and 29 stand rejected under 35 U.S.C. § 112, first paragraph, for alleged lack of adequate written description of the invention. The Examiner's argument on this issue is that the specification fails to describe any particular amino acid sequence that defines a protein as having raffinose synthase activity, and therefore the generic invention is not described.

Appellants submit that this argument is not persuasive. In the first instance, the specification asserts that the defined sequences in SEQ ID Nos: 1-8 (of which SEQ ID Nos:3-8 are recited in claims) define nucleic acids according to the invention, either at the nucleic acid or at the amino acid level. Appellants submit that specific description of a structure constitutes substantial evidence that they "possess" the invention so described and have placed such an invention in the hands of the public. *Vas-Cath v. Mahurkar*, 19 USPQ2d 1111 (Fed. Cir. 1991).

Furthermore, the specification describes a number of PCR primers, derived from the data of SEQ ID Nos: 2, 4, 6 and 8 or otherwise, that are useful when applied to template nucleic acids from plant types associated with the primer sequences as described in the specification, to obtain further cloned cDNAs encoding raffinose synthase enzymes. The specification also describes how to test any nucleic acids obtained by such a technique for activity of a raffinose synthase. Therefore, the invention is at the very least well-described in "product-by-process" terms. *Fiers v. Revel*, 25 USPQ2d at 1605. One may also consider that the PCR primers represent minimal nucleotide sequences that must be present to define a nucleic acid as one encoding a raffinose synthase. Also, the specification, at pages 20-21, describes particular regions of amino acid sequence that should have high homology to SEQ ID NO: 3, which is an amino acid sequence shown by Declaration evidence to represent a protein having RFS activity. Therefore, to this degree at least, a "structure-function" relationship is described in the specification.

Thus, Appellants submit that the specification meets the legal standard for adequate written description of the claimed invention, *i.e.* it evidences that the inventors were in possession of the invention as claimed. Accordingly, the rejection of claims 1, 4, 5, 8-10, 16-23, 28 and 29 under 35 U.S.C. § 112, first paragraph, for alleged lack of written description support, should be reversed.

Claims 1, 4, 5, 8-10, 16-23, 28 and 29 are also are rejected under 35 U.S.C. § 112, first paragraph, for alleged lack of enablement. The Examiner's position is essentially that, since one of ordinary skill in the art is unable to distinguish a nucleic acid encoding a raffinose synthase enzyme from a nucleic acid encoding a stachyose synthase enzyme based only on a degree of

sequence identity, the specification fails to teach the skilled artisan how to use the present invention.

This rejection fails in the first instance because the Examiner fails to establish any *prima facie* lack of enablement. Proper consideration of the question of enablement requires establishing that undue experimentation is required to practice the full scope of the invention. This question is addressed by considering a number of factors. *In re Wands*, 8 USPQ2d at 1400.

However, the Examiner's explanation of the rejection addresses only the question of whether one of ordinary skill in the art, having a particular nucleic acid in hand, can predict, based upon its sequence, whether or not that nucleic acid encodes a raffinose synthase enzyme, or whether instead it encodes a stachyose synthase. Such analysis ignores the other factors to be considered.

On the other hand, Appellants explain that the specification is enabling of the claimed invention, addressing the remaining considerations required under *Wands*. Appellants also present evidence to support an allegation that the skilled artisan, using the teachings of the specification in a manner accepted in the art at the time the invention was made (*e.g.* molecular phylogeny based upon degree of amino acid sequence similarity) can easily distinguish a raffinose synthase enzyme from a stachyose synthase enzyme. Appellants also point out that the specification provides express guidance of how to determine biochemically if a protein expressed from a cloned nucleic acid exhibits activity of a raffinose synthase. Furthermore, as to claims 5, 9 and 10, directed to particular nucleic acids encoding raffinose synthase enzymes, the specification describes utilities for the cloned nucleic acids that are independent of whether they actually encode a functional enzyme. For these three claims, the Examiner's entire rationale for making the rejection fails. Therefore it is plainly established that the present specification is enabling of the claimed invention and so the rejection of claims 1, 4, 5, 8-10, 16-23, 28 and 29 under 35 U.S.C. § 112, first paragraph, for alleged lack of enablement, must be reversed.

The favorable action of reversal of all of the rejection of claims 1, 4, 5, 8-10, 16-23, 28 and 29 under 35 U.S.C. § 112, first paragraph, for alleged lack of written description support and

for alleged lack of enablement, and remand to the Examiner for allowance of all of the pending claims, is respectfully requested.

VIII. CLAIMS

A copy of the claims involved in the present appeal is attached hereto as Appendix A. The claims in Appendix A are as addressed by the Examiner in the Final Office Action of August 23, 2006.

IX. EVIDENCE

A copy of evidence pursuant to §§ 1.130, 1.131, or 1.132 and/or evidence entered by or relied upon by the examiner that is relevant to this appeal is attached hereto as Appendix B.

1. Tables 1 and 2 and Figure. 1, which were presented attached to Appellants' paper of February 11, 2004.

2. Watanabe Declaration, presented attached to Appellants' paper of February 11, 2004

3. Exhibit 1, explanation of various sequence analysis programs, attached to Appellants' paper of November 15, 2004.

4. Lehle and Tanner, *Eur. J. Biochem.* 38:103-110 (1973), cited at the bottom of page 31 of Specification.

5. Declaration of Akistu NAGASAWA, copied from the copending application 09/301,714 and attached to Appellants' paper of June 2, 2006.

6. Richmond et al., *Plant Physiol.* 124:495-498 (2000), cited by the Examiner in the Office Action of August 11, 2003.

7. Duggleby, *Gene* 190:245-249 (1997), cited by the Examiner in the Office Actions of February 6, 2002 and November 20, 2002.

8. Bowie et al., *Science* 247:1306-1310 (1990), cited by the Examiner in the Office Action of November 20, 2002.

9. Lazar et al., *Molecular, Cellular Biology* 8:1247-1252 (1988), cited by the Examiner in the Office Action of November 20, 2002.

10. Broun et al., *Science* 282:1315-1317 (1998), cited by the Examiner in the Office Action of November 20, 2002.

11. Peterbauer et al., *Planta* 215:839-846 (2002), cited by the Examiner in the Office Action of August 11, 2003 and December 2, 2005.

12. Exhibit 4, alignment of SEQ ID NO: 5 of the instant application with SEQ ID NO: 7 of the instant application.

X. RELATED PROCEEDINGS

There are no prior decisions of any Court or of the Board of Appeals and Interferences in this matter.

Dated: July 23, 2007

Respectfully submitted,

By_____
Mark J. Nuell
Registration No.: 36,623
BIRCH, STEWART, KOLASCH & BIRCH, LLP
8110 Gatehouse Road
Suite 100 East
P.O. Box 747
Falls Church, Virginia 22040-0747
(703) 205-8000
Attorney for Applicant

APPENDIX A

Claims Involved in the Appeal of Application Serial No. 09/301,766

The pending claims 1-10, 16-23, 28 and 29, are set forth below as amended on June 2, 2006. Claims 1, 4, 5, 8-10, 16-23, 28 and 29 are on appeal:

1. An isolated nucleic acid which comprises a polynucleotide encoding a protein that binds a D-galactosyl group through the α(1→6) bond to the hydroxyl group attached to the carbon atom at 6-position of the D-glucose residue in a sucrose molecule to form raffinose, wherein said polynucleotide comprises a nucleotide sequence selected from the group consisting of:

(a) a nucleotide sequence encoding the amino acid sequence as depicted in SEQ 1D NO: 3,

(b) a nucleotide sequence depicted by the 236th to 2584th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 4,

(c) a nucleotide sequence encoding the amino acid sequence as depicted in SEQ 1D NO: 5,

(d) a nucleotide sequence depicted by the 134th to 2467th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 6,

(e) a nucleotide sequence encoding the amino acid sequence as depicted in SEQ ID NO: 7,

(f) a nucleotide sequence depicted by the 1st to 1719th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 8,

(g) a nucleotide sequence obtained from a polynucleotide which is amplified from a nucleic acid obtained from beet with a combination of a PCR primer selected from the group consisting of SEQ ID NO: 11 and SEQ ID NO: 13 and a PCR

primer selected from the group consisting of SEQ ID NO: 12 and SEQ ID NO: 14, wherein said nucleotide sequence hybridizes with a nucleotide sequence complementary to the nucleotide sequence of (a) or (b), in a buffer comprising 0.9M NaCl and 0.09M citric acid at 65°C to 68°C, and

(h) a nucleotide sequence obtained from a polynucleotide which is amplified from a nucleic acid obtained from mustard or rapeseed with a combination of a PCR primer selected from the group consisting of SEQ ID NO: 15, SEQ ID NO: 17 and SEQ ID NO: 19 and a PCR primer selected from the group consisting of SEQ ID NO: 16, SEQ ID NO: 18 and SEQ ID NO: 20, wherein said nucleotide sequence hybridizes with a nucleotide sequence complementary to the nucleotide sequence of any one of (c) to (f), in a buffer comprising 0.9M NaCl and 0.09M citric acid at 65°C to 68°C.

4. An isolated nucleic acid comprising a nucleotide sequence encoding the amino acid sequence as depicted in SEQ ID NO: 3.

5. (Allowed) An isolated nucleic acid comprising the nucleotide sequence depicted by the 236th to 2584th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 4.

6. (Allowed) An isolated nucleic acid comprising a nucleotide sequence encoding the amino acid sequence as depicted in SEQ ID NO: 5.

7. An isolated nucleic acid comprising the nucleotide sequence depicted by the 134th to 2467th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 6.

8. An isolated nucleic acid comprising a nucleotide sequence encoding the amino acid sequence as depicted in SEQ ID NO: 7.

9. An isolated nucleic acid comprising the nucleotide sequence depicted by the 1st to

1719th nucleotides in the nucleotide sequence as depicted in SEQ ID NO: 8.

10. An isolated nucleic acid comprising the nucleotide sequence as depicted in SEQ ID NO: 4, SEQ ID NO: 6, or SEQ ID NO: 8.

16. An isolated nucleic acid comprising the nucleic acid of claim 1, which is operatively linked to a promoter.

17. A vector comprising the nucleic acid of claim 1.

18. A transformant, wherein the nucleic acid of claim 1 is introduced into a host cell.

19. A transformant, wherein the nucleic acid of claim 16 is introduced into a host cell.

20. A transformant, wherein the vector of claim 17 is introduced into a host cell.

21. The transformant of claim 18, wherein the host cell is a microorganism.

22. The transformant of claim 18, wherein the host cell is a plant cell.

23. A method for producing a raffinose synthase which comprises the steps of: culturing or growing the transformant of claim 18 to produce the raffinose synthase, and collecting the raffinose synthase.

28. The nucleic acid of claim 16, wherein said promoter is effective in a plant cell.

29. The nucleic acid of claim 16, wherein said promoter is effective in a yeast cell.

APPENDIX B

The following items are of record as evidence in the present application and are attached hereto in support of Appellants' Appeal Brief:

1. Tables 1 and 2 and Figure. 1, which were presented attached to Appellants' paper of February 11, 2004.

2. Watanabe Declaration, presented attached to Appellants' paper of February 11, 2004

3. Exhibit 1, explanation of various sequence analysis programs, attached to Appellants' paper of November 15, 2004.

4. Lehle and Tanner, *Eur. J. Biochem.* 38:103-110 (1973), cited at the bottom of page 31 of Specification.

5. Declaration of Akistu NAGASAWA, copied from the copending application 09/301,714 and attached to Appellants' paper of June 2, 2006.

6. Richmond et al., *Plant Physiol.* 124:495-498 (2000), cited by the Examiner in the Office Action of August 11, 2003.

7. Duggleby, *Gene* 190:245-249 (1997), cited by the Examiner in the Office Actions of February 6, 2002 and November 20, 2002.

8. Bowie et al., *Science* 247:1306-1310 (1990), cited by the Examiner in the Office Action of November 20, 2002.

9. Lazar et al., *Molecular, Cellular Biology* 8:1247-1252 (1988), cited by the Examiner in the Office Action of November 20, 2002.

10. Broun et al., *Science* 282:1315-1317 (1998), cited by the Examiner in the Office Action of November 20, 2002.

11. Peterbauer et al., *Planta* 215:839-846 (2002), cited by the Examiner in the Office Action of August 11, 2003 and December 2, 2005.

12. Exhibit 4, alignment of SEQ ID NO: 5 of the instant application with SEQ ID NO: 7 of the instant application, attached to the present Appeal Brief.

Table 1

Code	Protein*	Organism	Accession**	Reference	Author/Assignee
Sc-03	RFS	***Beta vulgaris***	E37133	09/301,766	Sumitomo Chemical
Sc-05	RFS	***Brassica juncea***	E36417	09/301,766	Sumitomo Chemical
Sc-02	RFS	***Vicia faba***	E24423	08/992,914	Sumitomo Chemical
Sc-04	RFS	***Glycine max***	E24424	08/992,914	Sumitomo Chemical
Aj-05	RFS	***Cucumis sativus***	AF073744	Family GH36***	Ohsumi et al.
PsRFS	RFS	***Pisum sativum***	AJ426475	Family GH36	Peterbauer et al.
HvSIP	SIP	***Hordeum vulgare***	M77475	Family GH36	Heck et al.
PsSTS-1	STS	***Pisum sativum***	AJ311087	Family GH36	Peterbauer et al.
PsSTS-2	STS	***Pisum sativum***	AJ512932	Family GH36	Peterbauer et al.
VaSTS	STS	***Vigna angularis***	Y19024	Family GH36	Peterbauer et al.
AmSTS	STS	***Alonsoa meridionalis***	AJ487030	Family GH36	Voitsekhovskaja
SsSTS	STS	***Stachys affinis***	AJ344091	Family GH36	Pesch and Schmitz

*Protein: RFS, Raffinose synthase; SIP, Seed Imbibition Protein; STS, Stachyose synthase.

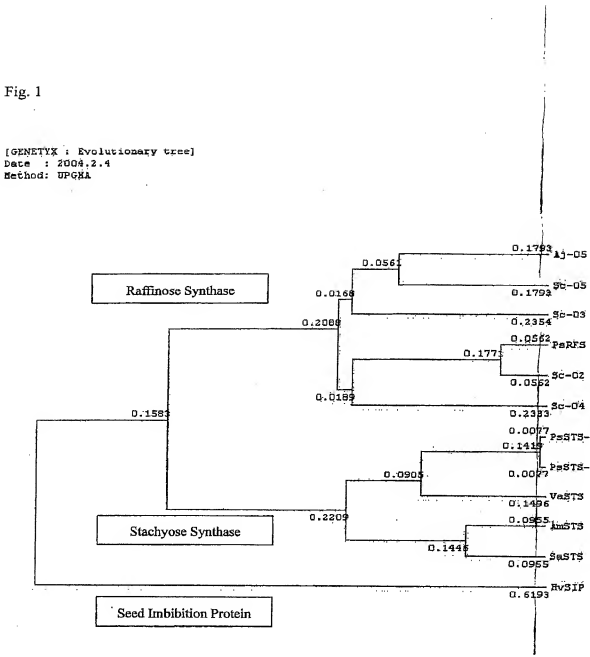
**Accession: GenBank Accession Number.

***Family GH36: glycoside hydrolase family 36 (see Carbohydrate-Active Enzymes (CAZy) database: http://afmb.cnrs-mrs.fr/CAZY/GH_36.html)

Table 2

Code	Sc-02	Sc-03	Sc-04	Sc-05	PsRFS	Aj-05	HvSIP	AmSTS	PsSTS-1	PsSTS-2	SaSTS	VaSTS
amino acids	800	783	611	777	798	784	757	868	853	847	863	857
Sc-02		62	54	62	89	64	39	43	43	42	42	43
Sc-03			50	63	63	66	39	42	43	42	42	43
Sc-04				50	54	50	29	35	35	35	34	36
Sc-05					62	70	39	44	44	43	44	45
PsRFS						64	38	42	43	43	42	43
Aj-05							38	42	44	43	43	43
HvSIP								33	34	34	34	34
AmSTS									64	64	83	64
PsSTS-1										98	65	75
PsSTS-2											65	74
SaSTS												65
VaSTS												

Fig. 1


[GENETYX : Evolutionary tree]
Date : 2004.2.4
Method: UPGMA
Raffinose Synthase
Stachyose Synthase
Seed Imbibition Protein
0.1793
0.0561
Aj-05
0.1793
Sc-05
0.0168
0.2354
Sc-03
0.2080
0.0562
PsRFS
0.1773
Sc-02
0.0562
0.0189
0.2333
Sc-04
0.1583
0.0077
PsSTS-1
0.1419
PsSTS-2
0.0077
0.0905
VaSTS
0.1496
0.2209
0.0955
AmSTS
0.1446
SaSTS
0.0955
RvSIP
0.6193

PATENT
Docket No. 20-4559P

IN THE U.S. PATENT AND TRADEMARK OFFICE

Applicants:	Eijiro WATANABE et al.		
Serial No.:	09/301,766	Group:	1638
Filed:	April 29, 1999	Examiner:	D.H.Kruse
For:	RAFFINOSE SYNTHASE GENES AND THEIR USE		

DECLARATION UNDER 37 CFR 1.132

Honorable Commissioner of Patents and Trademarks
Washington, D.C. 20231

Sir:

I, Eijiro WATANABE, citizen of Japan and residing in Fukui-cho 32-12-403, Takarazuka-shi, Hyogo-ken, Japan, declare and say that:

1. I completed the doctor's course, with a major in agricultural chemistry, of the graduate school of Tokyo University and obtained a doctor's degree in agriculture at Tokyo University in March, 1991.

2. From April, 1991, I made further researches in the Department of Agricultural Chemistry, Faculty of Agriculture, Tokyo University, as a postdoctoral fellow (Japan Society for the Promotion of Science) for one year.

3. From April, 1992 to the present, I have been an employee of Sumitomo Chemical Company, Limited, the assignee of the above-identified application.

4. From April, 1992 to March, 2000, I had been engaged in research works for plant engineering using recombination and other gene manipulation, such as cloning of plant genes, preparation and evaluation of transgenic plants.

5. I am one of the inventors of the above-identified application and am familiar with the subject matter thereof.

6. I have read the Office Action mailed August 11, 2003 and the reference cited, and am familiar with the subject matter thereof.

7. To demonstrate successful expression of raffinose synthase activity in transgenic plants, I have made the following experiments.

Experiments

Transformation of Tobacco with Raffinose Synthase Gene Derived from Brassica Plant

The vector BjRS-Sac(+)-121 having the mustard raffinose synthase gene of the present invention in the expressible direction (*i.e.* sense direction) and the vector BjRS-Sac(-)-121 having the mustard raffinose synthase gene of the present invention in the reverse direction (*i.e.* antisense direction), which are the same as obtained in Example 8 of the present specification, were used for the transformation of tobacco (*Nicotiana tubacum*) by the Agrobacterium infection method.

Agrobacterium tumefaciens (strain LBA4404 having rifampicin and streptomycin resistance) previously converted into a competent state by calcium chloride treatment was transformed independently with two plasmids BjRS-Sac(+)-121 and BjRS-Sac(-)-121. The transformants were selected on LB medium containing 50 μg/ml rifampicin and 25 μg/ml kanamycin by utilizing the kanamycin resistant character conferred by the kanamycin resistant gene (neomycin phosphotransferase, NPTII) of the introduced plasmids.

The transformant Agrobacterium obtained (*Agrobacterium tumefaciens* strain LBA4404, rifampicin and streptomycin resistant) was cultured on LB medium containing 50 μg/ml rifampicin and 25 μg/ml kanamycin at 28°C for a whole day and night, and the culture was used for the transformation of tobacco by the method described below.

Seeds of tobacco were aseptically sown on 1/2 MS medium containing 2% sucrose and 0.7% agar. After one week, leaves of sprouting plants were cut out with a scalpel, and transferred to MS medium containing 3% sucrose, 0.7% agar, 1.0mg/l BA and 0.1mg/l NAA, followed by preculture for 1 day. The precultured leaves were transferred in a 1000-fold dilution of the Agrobacterium culture broth and allowed to stand for 5 minutes. The leaves were transferred again to the same medium as used in the preculture, and cultured for 3 to 4 days. The cultured leaves were transferred to MS medium containing 3% sucrose, 1.0mg/l BA, 0.1mg/l NAA and 500 mg/l cefotaxim, and shaken for 1 day to remove microbial cells. The leaves thus treated were transferred to

MS medium containing 3% sucrose, 0.7% agar, 1.0mg/l BA, 0.1mg/l NAA, 100 mg/l cefotaxim and 20 mg/l kanamycin, and cultured for 3 to 4 weeks. The leaves were transferred to MS medium containing 3% sucrose, 0.7% agar, 1.0mg/l BA, 0.1mg/l NAA, 100 mg/l cefotaxim and 20 mg/l kanamycin, and cultivated. The cultivation on this medium was continued with subculturing at intervals of 3 to 4 weeks. When shoots were began to regenerate, these shoots were subcultured on MS medium containing 3% sucrose, 0.7% agar and 20 mg/l kanamycin, and cultivated for 3 to 4 weeks. The rooting plants were transferred to vermiculite : peat moss = 1 : 1, and cultivated at 21°C to 22°C in a cycle of day/night = 12 hours : 12 hours. With the progress of plant body growth, the plants were grown with cultivation soil.

Measurement of Raffinose Synthase Activity

Leaves of the transformed tobacco plant were put in 10 times of the leaf weight of 100 mM Tris-HCl (pH 7.4), 5 mM DTT (dithiothreitol), 1 mM EDTA, 1 mM PMSF (phenylmethylsulfonyl fluoride) and 1 mM benzamide, and ground on ice with a mortar. The ground material was centrifuged at 21,400 x g for 50 minutes at 4°C. The resulting supernatant was recovered and used as a sample for the following measurement of raffinose synthase activity.

The raffinose synthase activity was measured under the following conditions according to the description of L. Lehle and W. Tanner, Eur. J. Biochem., 38, 103-110 (1973).

First, 2 μl of a sample to be used in the measurement of activity was added to 18 μl of the reaction mixture that came to contain 100 mM Tris-HCl (pH 7.4), 5 mM DTT (dithiothreitol), 0.01% BSA, 200 μM sucrose, 5 mM galactinol, 31.7 μM [^{14}C] sucrose, and the reaction mixture was kept at 37°C for 18.3 hours. After the reaction, 30 μl of ethanol was added to the reaction mixture, followed by stirring and centrifugation at 15,000 rpm for 5 minutes. The supernatant was spotted at a volume of 5 μl on an HPTLC plate of cellulose for thin layer chromatography (Merck, 10 cm x 20 cm), and developed with n-butanol : pyridine : water : acetic acid = 60 : 40 : 30 : 3. The developed plate was dried and then quantitatively analyzed with an imaging analyzer (Fuji Photographic Film, FUJIX Bio Imaging Analyzer BAS-2000II) for the

determination of content of [^{14}C] raffinose. Raffinose synthase activity in each sample was calculated from the content of [^{14}C] raffinose.

Results

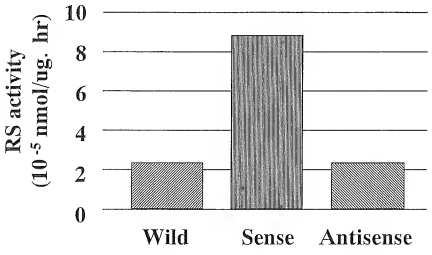
Results are summarized in Fig. 1. The transformed tobacco plant with BjRS-Sac(+)-121 ("Sense" in the figure) showed significantly higher level of raffinose synthase activity in leaf than the wild type ("Wild" in the figure).

Discussion

As can be seen from Fig. 1, the transformed tobacco plant having the mustard raffinose synthase gene of the present invention in sense direction exhibited higher raffinose synthase activity as compared with the control tobacco plant having no such gene. This indicates that tabacco plants may have improved raffinose synthase activity by introduction of the raffinose synthase gene of the present invention in sense direction into these plants.

Thus, it is clearly demonstrated that the raffinose synthase gene of the present invention can successfully express raffinose synthase activity in the transformed tobacco plant.

Fig.1



8. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonments, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the above-identified application or any patent issued thereon.

This 6th day of February, 2004

E. Watanabe

Eijiro WATANABE

IN THE U.S. PATENT AND TRADEMARK OFFICE

Applicants:	Eijiro WATANABE et al.		
Serial No.:	08/992,914	Group:	1638
Filed:	December 18, 1997	Examiner:	D.H.Kruse
For:	RAFFINOSE SYNTHASE GENES AND THEIR USE		

DECLARATION UNDER 37 CFR 1.132

Honorable Commissioner of Patents and Trademarks
Washington, D.C. 20231

Sir:

I, Akitsu NAGASAWA, citizen of Japan and residing in Kamokogahara 3-28-56, Higashi-Nada-ku, Kobe-shi, Hyogo-ken, Japan, declare and say that:

1. I completed the master's course, with a major in agricultural biology, of the graduate school of Kyoto University and obtained a master's degree in agriculture at Kyoto University in March, 1984.

2. From April, 1984 to the present, I have been an employee of Sumitomo Chemical Company, Limited, the assignee of the above-identified application.

3. From April, 1984 to the present, I have been engaged in research works for plant engineering using recombination and other gene manipulation, such as cloning of plant genes, preparation and evaluation of transgenic plants.

4. I am one of the members of the research project related to the above-identified application and am familiar with the subject matter thereof.

5. I have read the Office Action mailed March 11, 2005 and the reference cited, and am familiar with the subject matter thereof.

6. To demonstrate successful identification of raffinose synthase genes in plant, I have made the following computer analysis.

ANALYSIS

1) The overall sequence homologies (%) among the amino acid sequences of raffinose synthases (RFSs), seed imbibition protein (SIP) and stachyose synthases (STSs) shown in Table 1 attached hereto were calculated based on a global multiple alignment (the alignment of sequences over their entire length) using the gene analysis software GENETYX-SV/RC for Windows version 6.1.0 (GENETYX Corporation; http://www.sdc.co.jp/genetyx/) with default parameters. The global multiple alignment was generated using CLUSTAL sequence analysis program. The amino acid sequences of the RFSs, SIP and STSs used to produce the global multiple alignment are as follows:

Sc-02:
MAPPSITKTATLQDVISTIDIGNGNSPLFSITLDQSRDFLANGHPFLTQV
PPNITTTTTTTASSFLNLKSNKDTIPNNNNTMLLQQGCFVGFNSTEPKSH
HVVPLGKLKGIKFMSIFRFKVWWTTHWVGTNGQELQHETQMLILDKNDSL
GRPYVLLLPILENTFRTSLQPGLNDHIGMSVESGSTHVTGSSFKACLYIH
LSNDPYSILKEAVKVIQTQLGTFKTLEEKTAPSIIDKFGWCTWDAFYLKV
HPKGVWEGVKSLTDGGCPPGFVIIDDGWQSICHDDDDEDDSGMNRTSAGE
QMPCRLVKYEENSKFREYENPENGGKKGLGGFVRDLKEEFGSVESVYVWH
ALCGYWGGVRPGVHGMPKARVVVPKVSQGLKMTMEDLAVDKIVENGVGLV
PPDFAHEMFDGLHSHLESAGIDGVKVDVIHLLELLSEEYGGRVELARAYY
KALTSSVKKHFKGNGVIASMEHCNDFFLLGTEAISLGRVGDDFWCSDPSG
DPNGTYWLQGCHMVHCAYNSLWMGNFIQPDWDMFQSTHPCAEFHAASRAI
SGGPIYVSDCVGNHNFKLLKSLVLPDGSILRCQHYALPTRDCLFEDPLHN
GKTMLKIWNLNKYTGVLGLFNCQGGGWCPEARRNKSVSEFSRAVTCYASP
EDIEWCNGKTPMSTKGVDFFAVYFFKEKKLRLMKCSDRLKVSLEPFSFEL
MTVSPVKVFSKRFIQFAPIGLVNMLNSGGAIQSLEFDDNASLVKIGVRGC
GEMSVFASEKPVCCKIDGVKVKFLYEDKMARVQILWPSSSTLSLVQFLF

Sc-03:
MAPSFSKENSKTCDEVANHDDCNTCPIISLEESNFMVNGHVILSQVPSNI
TAISKMGFDGLFVGFDAPEPKARHWVSVGQLKGIPFMSIFRFKVWWTTHW
TGSNGRDLEHETQILILDKSDEGLGRPYIVILPLIEGPFRASLQPGSVDD
YVDICVESGSTKVVGDSFRAVLYIRAGPDPFKLIKDTMKEVQAHLGTFKL
LDDKTPPGIVDKFGWCTWDAFYLKVEXYGVWEGVKGLVENGVPPGLVLID
DGWQSICHDDDPITDQEGINRTSAGEQMPCRLIKYEENFKFRDYKSPNIM
GHEDHPNMGMRAFVRDLKEEFKTVEHVYVWHAFTGYWGGVRPNVPGLXEA
QVVTPKLSPGLEMTMEDLAVDKIVNNGIGLVQPDKAQELYEGLHSHLENC

GIDGVKVDVIHLLEMMAEDYGGRVELAKTYYKAITESVRKHFKGNGVIAS
MEQCNDFMLLGTETICLGRVGDDFWPTDPSGDINGTYWLQGCHMVHCAYN
SLWMGNFIHPDWDMFQSTHPCAEFHAASRAISGGPIYVSDVVGKHNIPLL
KRLVLADGSILRCEYHALPTKDCLFVDPLHDGKTMLKIWNLNKYNGVLGV
FNCQGGGWSRESRKNLCFSEYSKPISCKTSPKDVEWENGHKPFPIKGVEC
FAMYFTKEKKLILSQLSDTIEISLDPFDYELIVVSPMTILPWESIAFAPI
GLVNMLNAGGAVKSLDISEDNEDKMVQVGIKGAGEMMVYSSEKPKACRVN
GEDMEFEYEESMIKVQVTWNHNSGGFTTVEYLF

Sc-04 (truncated):
MAPSISKTVELNSFGLVNGNLPLSITLEGSNFLANGHPFLTEVPENIIVT
PSPIDAKSSKNNEDDDVVGCFVGFHADEPRSRHVASLGKLRGIKFMSIFR
FKVWWTTHWVGSNGHELEHETQMMLLDKNDQLGRPFVLILPILQASFRAS
LQPGLDDYVDVCMESGSTRVCGSSFGSCLYVHVGHDPYQLLREATKVVRM
HLGTFKLLEEKTAPVIIDKFGWCTWDAFYLKVHPSGVWEGVKGLVEGGCP
PGMVLIDDGWQAICHDEDPITDQEGMKRTSAGEQMPCRLVKLEENYKFRQ
YCSGKDSEKGMGAFVRDLKEQFRSVEQVYVWHALCGYWGGVRPKVPGMPQ
AKVVTPKLSNGLKLTMKDLAVDKIVSNGVGLVPPHLAHLLYEGLHSRLES
AGIDGVKVDVIHLLEMLSEEYGGRVELAKAYYKALTASVKKHFKGNGVIA
SMEHCNDFPLLGTEAIALGRVGDDFWCTDPSGDPNGTYWLQGCHMVHCAY
NSLWMGNFIQPDWDMFQSTHPCAEFHAPLGPSLVDQFTLVIVLESTTSSC
SRASLCLMGRFCVVNTMHSPHETVCLKTPCMMGRQCSKFGISTNIQVFWV
YLIAKEVGGVP

Sc-05:
MAPPSVIKSDAAVNGIDLSGKPLFRLEGSDLLANGHVVLTDVPVNVTVTA
SPYLADKDGEPVDASAGSFIGFNLDGEPRSRHVASIGKLRDIRFMSIFRF
KVWWTTHWVGSKGSDIENETQIIILENSGSGRPYVLLLPLLEGSFRSSFQ
PGEDDDVAVCVESGSTQVTGSEFRQVVYVHAGDDPFKLVKDAMKVVRVHM
NTFKLLEEKXPPGIVDKFGWCTWDAFYLTVNPDGVHKGVKCLVDGGCPPG
LVLIDDGWQSIGHDSDGIDVEGMSCTVAGEQMPCRLLKFQENFKFRDYVS
PKDKNEVGMKAFVRDLKEEFSTVDYIYVWHALCGYWGGLRPGAPTLPPST
IVRPELSPGLKLTMQDLAVDKIVDTGIGFVSPDMANEFYEGLIISHLQNVG
IDGVKVDVIHILEMLCEKYGGRVDLAKAYFKALTSSVNKHFDGNGVIASM
EHCNDFMFLGTEAISLGRVGDDFWCTDPSGDINGTYWLQGCHMVHCAYNS
LWMGNFIQPDWDMFQSTHPCAEFHAASRAISGGPIYISDCVGQHDFDLLK
RLVLPDGSILRCEHYALPTRDRLFEDPLHDGKTMLKIWNLNKYTGIIGAF
NCQGGGWCRETRRNQCFSQCVNTLTATTNPKDVEWNSGNNPISVENVEEF
ALFLSQSKKLVLSGPNDDLEITLEPFKFELITVSPVVTIEGSSVQFAPIG
LVNMLNTSGAIRSLVYHEESVEIGVRGAGEFRVYASRKPASCKIDGEVVE
FGYEESMVMVQVPWSAPEGLSSIKYEF

PsRFS:
MAPPSITKTATQQDVISTVDIGNSPLLSISLDQSRNFLVNGHPFLTQVPP
NITTTTTSTPSPFLDFKSNKDTIANNNNTLQQQGCFVGFNTTEAKSHHVV
PLGKLKGIKFTSIFRFKVWWTTHWVGTNGHELQHETQILILDKNISLGRP
YVLLLPILENSFRTSLQPGLNDYVDMSVESGSTHVTGSTFKACLYLHLSN
DPYRLVKEAVKVIQTKLGTFKTLEEKTPPSIIEKFGWCTWDAFYLKVHPK
GVWEGVKALTDGGCPPGFVIIDDGWQSISHDDDDPVTERDGMNRTSAGEQ
MPCRLIKYEENYKFREYENGDNGGKKGLVGFVRDLKEEFRSVESVYVWHA
LCGYWGGVRPKVCGMPEAKVVVPKLSPGVKMTMEDLAVDKIVENGVGLVP
PNLAQEMFDGIHSHLESAGIDGVKVDVIHLLELLSEEYGGRVELAKAYYK
ALTSSVNKHFKGNGVIASMEHCNDFFLLGTEAISLGRVGDDFWCCDPSGD
PNGTYWLQGCHMVHCAYNSLWMGNFIHPDWDMFQSTHPCAEFHAASRAIS
GGPVYVSDCVGNHNFKLLKSFVLPDGSILRCQHYALPTRDCLFEDPLHNG
KTMLKIWNLNKYAGVLGLFNCQGGGWCPETRRNKSASEFSHAVTCYASPE
DIEWCNGKTPMDIKGVDVFAVYFFKEKKLSLMKCSDRLEVSLEPFSFELM
TVSPLKVFSKRLIQFAPIGLVNMLNSGGAVQSLEFDDSASLVKIGVRGCG
ELSVFASEKPVCCKIDGVSVEFDYEDKMVRVQILWPGSSTLSLVEFLF

Aj-05:
MAPSFKNGGSNVVSFDGLNDMSSPFAIDGSDFTVNGHSFLSDVPENIVAS
PSPYTSIDKSPVSVGCFVGFDASEPDSRHVVSIGKLKDIRFMSIFRFKVW
WTTHWVGRNGGDLESETQIVILEKSDSGRPYVFLLPIVEGPFRTSIQPGD
DDFVDVCVESGSSKVVDASFRSMLYLHAGDDPFALVKEAMKIVRTHLGTF
RLLEEKTPPGIVDKFGWCTWDAFYLTVHPQGVIEGVRHLVDGGCPPGLVL
IDDGWQSIGHDSDPITKEGMNQTVAGEQMPCRLLKFQENYKFRDYVNPKA
TGPRAGQKGMKAFIDELKGEFKTVEHVYVWHALCGYWGGLRPQVPGLPEA
RVIQPVLSPGLQMTMEDLAVDKIVLHKVGLVPPEKAEEMYEGLHAHLEKV
GIDGVKIDVIHLLEMLCEDYGGRVDLAKAYYKAMTKSINKHFKGNGVIAS
MEHCNDFMFLGTEAISLGRVGDDFWCTDPSGDPNGTFWLQGCHMVHCAND
SLWMGNFIHPDWDMFQSTHPCAAFHAASRAISGGPIYVSDSVGKHNFDLL
KKLVLPDGSILRSEYYALPTRDCLFEDPLHNGETMLKIWNLNKFTGVIGA
FNCQGGGWCRETRRNQCFSQYSKRVTSKTNPKDIEWHSGENPISIEGVKT
FALYLYQAKKLILSKPSQDLDIALDPFEFELITVSPVTKLIQTSLHFAPI
GLVNMLNTSGAIQSVDYDDDLSSVEIGVKGCGEMRVFASKKPRACRIDGE
DVGFKYDQDQMVVVQVPWPIDSSSGGISVIEYLF

IlvSIP:
MTVTPQITVGDGRLAVRGRTVLSGVPDNVTAAHAAGAGLVDGAFVGATAA
EAKSHHVFTFGTLRDCRFMCLFRFKLWWMTQRMGTSGRDVPLETQFILIE

VPAAAGNDDGDSSDGDSEPVYLVMLPLLEGQFRTVLQGNDQDELQICIES
GDKAVETEQGMNNVYVHAGTNPFDTITQAVKAVEKHTQTFHHREKKTVPS
FVDWFGWCTWDAFYTDVTADGVKQGLRSLAEGGAPPRFLIIDDGWQQIGS
ENKDDPGVAVQEGAQFASRLTGIRENTKFQSEHNQEETPGLKRLVDETKK
EHGVKSVYVWHAMAGYWGGVKPSAAGMEHYEPALAYPVQSPGVTGNQPDI
VMDSLSVLGLGLVHPRRVHRFYDELHAYLAACGVDGVKVDVQNIVETLGA
GHGGRVALTRAYHRALEASVARNFPDNGCISCMCHNTDMLYSAKQTAVVR
ASDDFYPRDPASHTVHISSVAYNTLFLGEFMQPDWDMFHSLHPAAEYHGA
ARAIGGCPIYVSDKPGNHNFDLLRKLVLPDGSVLRAQLPGRPTRDCLFSD
PARDGASLLKIWNMNKCAGVVGVFNCQGAGWCRVAKKTRIHDEAPGTLTG
SVRAEDVEAIAQAAGTGDWGGEAVVYAHRAGELVRLPRGATLPVTLKRLE
YELFHVCPVRAVAPGVSFAPIGLLHMFNAGGAVEECTVETGEDGNAVVGL
RVRGCGRFGAYCSRRPAKCSVDSADVEFTYDSDTGLVTADVPVPEKEMYR
CALEIRV

AmSTS:
MAPPYDPIPIPIPMSAILNFLSSTVKDNSPELLDGTLSVKNVPILTDIPS
NVSFSSFSSIVQSSEAPVPLFQRAQSLSSSGGFLGFSQNEPSSRLMNSLG
KFTDRDFVSIFRFKTWWSTQWVGTTGSDIQMETQWIMLDVPEIKSYAVVV
PIVEGKFRSALFPGKDGHILIGAESGSTKVKTSNFDAIAYVHVSENPYTL
MRDAYTAVRVHLNTFKLIEEKSAPPLVNKFGWWTWDAFYLTVEPAGIYHG
VQEFADGGLTPRFLIIDDGWQSINNDDNDPNEDAKNLVLGGTQMTARLHR
LDECEKFRKYKGGSMSGPNRPPFDPKKPKLLISKAIEIEVAEKARDKAAQ
SGVTDLARYEAEIEKLTKELDQMFGGGGEETSSGKSCSSCSKSDNFGMK
AFTKDLRTNFKGLDDIYVWHALAGAWGGVRPGATHLNAKIVPTNLSPGLD
GTMTDLAVVKIIEGSTGLVDPDQAEDFYDSMHSYLSSVGITGVKVDVIHT
LEYISEDYGGRVELAKAYYKGLSKSLAKNFNGTGLISSMQQCNDFFLLGT
EQISMGRVGDDFWFQDPNGDPMGVYWLQGVHMIHCAYNSMWMGQFIQPDW
DMFQSDHPGGYFHAGSRAICGGPVYVSDSLGGHNFDLLKKLVFNDGTIPK
CIHFALPTRDCLFKNPLFDSKTILKIWNFNKYGGVIGAFNCQGAGWDPKE
QRIKGYSQCYKPLSGSVHVSGIEFDQKKEASEMGEAEEYAVYLSEAEKLS
LATRDSDPIKITIQSSTFEIFSFVPIKKLGEGVKFAPIGLTNLFNAGGTI
QGLVYNEGIAKIEVKGDGKFLAYSSVVPKKAYVNGAEKVFAWSGNGKLEL
DITWYEECGGISNVTFVY

PsSTS-1:
MAPPLNSTTSNLIKTESIFDLSERKFKVKGFPLFHDVPENVSFRSFSSIC
KPSESNAPPSLLQKVLAYSHKGGFFGFSHETPSDRLMNSIGSFNGKDFLS
IFRFKTWWSTQWIGKSGSDLQMETQWILIEVPETKSYVVIIPIIEKCFRS
ALFPGFNDHVKIIAESGSTKVKESTFNSIAYVHFSENPYDLMKEAYSAIR
VHLNSFRLLEEKTIPNLVDKFGWCTWDAFYLTVNPIGIFHGLDDFSKGGV

EPRFVIIDDGWQSISFDGYDPNEDAKNLVLGGEQMSGRLHRFDECYKFRK
YESGLLLGPNSPPYDPNNFTDLILKGIEHEKLRKKREEAISSKSSDLAEI
ESKIKKVVKEIDDLFGGEQFSSGEKSEMKSEYGLKAFTKDLRTKFKGLDD
VYVWHALCGAWGGVRPETTHLDTKIVPCKLSPGLDGTMEDLAVVEISKAS
LGLVHPSQANELYDSMHSYLAESGITGVKVDVIHSLEYVCDEYGGRVDLA
KVYYEGLTKSIVKNFNGNGMIASMQHCNDFFFLGTKQISMGRVGDDFWFQ
DPNGDPMGSFWLQGVHMIHCSYNSLWMGQMIQPDWDMFQSDHVCAKFHAG
SRAICGGPIYVSDNVGSHDFDLIKKLVFPDGTIPKCIYFPLPTRDCLFKN
PLFDHTTVLKIWNFNKYGGVIGAFNCQGAGWDPIMQKFRGFPECYKPIPG
TVHVTEVEWDQKEETSHLGKAEEYVVYLNQAEELSLMTLKSEPIQFTIQP
STFELYSFVPVTKLCGGIKFAPIGLTNMFNSGGTVIDLEYVGNGAKIKVK
GGGSFLAYSSESPKKFQLNGCEVDFEWLGDGKLCVNVPWIEEACGVSDME
IFF

PsSTS-2:
MAPPLNSTTSNLIKTESIFDLSERKFKVKGFPLFHDVPENVSFRSFSSIC
KPSESNAPPSLLQKVLAYSHKGGFFGFSHETPSDRLMNSLGSFNGKDFLS
IFRFKTWWSTQWIGKSGSDLQMETQWILIEVPETKSYVVIPIIEKCFRS
ALFPGFNDHVKIIAESGSTKVKESTFNSIAYVHFSENPYDLMKEAYIAIR
VHLNSFRLLEEKTIPNLVDKFGWCTWDAFYLTVNPIGIFHGLDDFSKGGV
EPRFVIIDDGWQSISFDGCDPNEDAKNLVLGGEQMSGRLHRFDECYKFRK
YESGLLLGPNSPPYDPKKFTDLILKGIEHEKLRKKREEAISSKSSDLAEI
ESKIKKVVKEIDDLFGGEQFSSVEKSEMKSEYGLKAFTKDLRTKFKGLDD
VYVWHALCGAWGGVRPETTHLDTKFVPCKLSPGLDGTMEDLAVVEISKAS
LGLVHPSQANELYDSMHSYLAESGITGVKVDVIHSLEYVCDEYGGRVDLA
KVYYEGLTKSIVKNFNGNGMIASMQQCNDFFFLGTKQISMGRVGDDFWFQ
DPNGDPMGSFWLQGVHMIHCSYNSLWMGQMIQPDWDMFKSDHVCAKFHAG
SRAICGGPIYVSDNVGSHDFDLIKKLVFPDGTIPKCIYFPLPTRDCLFKN
PLFDHTTLLKIWNFNKYGGVIGAFNCQGAGWDPIMQKFRGFPECYKPIPG
TVHVTQVEWDQKEETSHFGKAEEYVVYLNQAEELCLMTLKSEPIQFTIQP
STFELYSFVPVTKLCGGIKFAPIGLTNMFNSGGTVIDLEYVGNGAKIKVK
GGGSFLAYSSESPKKFQLNGCEVDFEWLGDGKLCVNVPWIEEACGVS

SaSTS:
MAPPNDPISSIFSPLISVKKDNAFELVGGKLSVKNVPLLSEIPSNVTFKS
FSSICQSSGAPAPLYNRAQSLSNCGGFLGFSQKESADSVTNSLGKFTNRE
FVSIFRFKTWWSTQWVGTSGSDIQMETQWIMLNLPEIKSYAVVIPIVEGK
FRSALFPGKDGHVLISAESGSTCVKTTSFTSIAYVHVSDNPYTLMKDGYT
AVRVHLDTFKLIEEKSAPPLVNKFGWCTWDAFYLTVEPAGIWNGVKEFSD
GGFSPRFLIIDDGWQSINIDGQDPNEDAKNLVLGGTQMTARLHRFDECEK
FRKYKGGSMMGPKVPYFDPKKPKLLISKAIEIEGVEKARDKAIQSGITDL

SQYEIKLKKLNKELDEMFGGGGNDEKGSSKGCSDCSCKSQNSGMKAFTND
LRTNFKGLDDIYVWHALAGAWGGVKPGATHLNAKIEPCKLSPGLDGTMTD
LAVVKILEGSIGLVHPDQAEDFYDSMHSYLSKVGITGVKVDVIHTLEYVS
ENYGGRVELGKAYYKGLSKSLKKNFNGSGLISSMQQCNDFFLLGTEQISM
GRVGDDFWFQDPNGDPMGVFWLQGVHMIHCAYNSMWMGQIIHPDWDMFQS
DHCSAKFHAGSRAICGGPVYVSDSLGGHDFDLLKKLVFNDGTIPKCIHFA
LPTRDCLFKNPLFDSKTILKIWNFNKYGGVVGAFNCQGAGWDPKEQRIKG
YSECYKPLSGSVHVSDIEWDQKVEATKMGEAEEYAVYLTESEKLLLTTPE
SDPIPFTLKSTTFEIFSFVPIKKLGQGVKFAPIGLTNLFNSGGTIQGVVY
DEGVAKIEVKGDGKFLAYSSSVPKRSYLNGEEVEYKWSGNGKVEVDVPWY
EECGGISNITFVF

VaSTS:
MAPPNDPVNATLGLEPSEKVFDLSDGKLTVKGVVLLSHVPENVTFSSFSS
ICVPRDAPSSILQRVTAASHKGGFLGFSHVSPSDRLINSLGSFRGRNFLS
IFRFKTWWSTQWVGNSGSDLQMETQWILIEVPETESYVVIIPIIEKSFRS
ALHPGSDDHVKICAESGSTQVRASSFGAIAYVHVAETPYNLMREAYSALR
VHLDSFRLLEEKTVPRIVDKFGWCTWDAFYLTVNPVGVWHGLKDFSEGGV
APRFVVIDDGWQSVNFDDEDPNEDAKNLVLGGEQMTARLHRFEEGDKFRK
YQKGLLLGPNAPSFNPETIKELISKGIEAEHLGKQAAAISAGGSDLAEIE
LMIVKVREEIDDLFGGKGKESNESGGCCCKAAECGGMKDFTTDLRTEFKG
LDDVYVWHALCGGWGGVRPGTTHLDSKIIPCKLSPGLVGTMKDLAVDKIV
EGSIGLVHPHQANDLYDSMHSYLAQTGVTGVKIDVIHSLEYVCEEYGGRV
EIAKAYYDGLTNSIIKNFNGSGIIASMQQCNDFFFLGTKQIPFGRVGDDF
WFQDPNGDPMGVFWLQGVHMIHCSYNSLWMGQIIQPDWDMFQSDHECAKF
HAGSRAICGGPVYVSDSVGSHDFDLIKKLVFPDGTVPKCIYFPLPTRDCL
FRNPLFDQKTVLKIWNFNKYGGVIGAFNCQGAGWDPKGKKFKGFPECYKA
ISCTVHVTEVEWDQKKEAEHMGKAEEYVVYLNQAEVLHLMTPVSEPLQLT
IQPSTFELYNFVPVEKLGSSNIKFAPIGLTNMFNSGGTIQELEYIEKDVK
VKVKGGGRFLAYSTQSPKKFQLNGSDAAFQWLPDGKLTLNLAWIEENDGV
SDLAIFF

The calculated overall sequence homologies (%) are shown in Table 2 attached hereto. The homologies between RFSs and SIP are less than 40%. The homologies between RFSs and STSs are not higher than 45%. On the other hand, the homologies among RFSs are all 50% or higher. Thus, the homologies among RFSs are higher than those homologies between RFSs and SIP and between RFSs and STSs.

A molecular phylogenic tree of the RFSs, SIP and STSs shown in Table 1 is

drawn in Figure 1 attached hereto. The molecular phylogenic tree is drawn by the UPGMA method using the gene analysis software GENETYX-SV/RC for Windows version 6.1.0 (GENETYX Corporation; http://www.sdc.co.jp/genetyx/) with default parameters. In the molecular phylogenic tree, RFSs, SIP and STSs form different groups respectively.

In summary, Table 2 and Figure 1 show that RFSs, SIP and STSs can be distinguished from one another based upon a comparison of their amino acid sequences.

2) Attached Table 3 shows the identities obtained using the BLAST program for the amino acid sequences of RFSs, SIP and STSs shown in Table 1. Among Sc-02, Sc-03, Sc-04 and Sc-05, the identities were obtained by searching the "patent database" provided by NCBI (National Center for Biotechnology Information) with default parameters, using the amino acid sequence of each protein as the "query", and using "Protein query vs. translated database (tblastn)" of the NCBI BLAST program. Also, other identities were obtained by searching the "non-redundant database" provided by NCBI with default parameters, using the amino acid sequence of each protein as the "query", and using "Protein-protein BLAST (blastp)" of the NCBI BLAST program. The above-identified amino acid sequences of the RFSs, SIP and STSs are used as the "query" except that the amino acid sequence of Sc-04 used as the "query" is as follows:

Sc-04 (full-length):
MAPSISKTVELNSFGLVNGNLPLSITLEGSNFLANGHPFLTEVPENIIVT
PSPIDAKSSKNNEDDDVVGCFVGFHADEPRSRHVASLGKLRGIKFMSIFR
FKVWWTTHWVGSNGHELEHETQMMLLDKNDQLGRPFVLILPILQASFRAS
LQPGLDDYVDVCMESGSTRVCGSSFGSCLYVHVGHDPYQLLREATKVVRM
HLGTFKLLEEKTAPVIIDKFGWCTWDAFYLKVHPSGVWEGVKGLVEGGCP
PGMVLIDDGWQAICHDEDPITDQEGMKRTSAGEQMPCRLVKLEENYKFRQ
YCSGKDSEKGMGAFVRDLKEQFRSVEQVYVVWHALCGYWGGVRPKVPGMPQ
AKVVTPKLSNGLKLTMKDLAVDKIVSNGVGLVPPHLAHLLYEGLHSRLES
AGIDGVKVDVIHLLEMLSEEYGGRVELAKAYYKALTASVKKHFKGNGVIA
SMEHCNDFFLLGTEAIALGRVGDDFWCTDPSGDPNGTYWLQGCHMVHCAY
NSLWMGNFIQPDWDMFQSTHPCAEFHAASRAISGGPVYVSDCVGKHNFKL
LKSLALPDGTILRCQHYALPTRDCLFEDPLHDGKTMLKIWNLNKYTGVLG

LFNCQGGGWCPVTRRNKSASEFSQTVTCLASPQDIEWSNGKSPICIKGMN
VFAVYLFKDHKLKLMKASEKLEVSLEPFTFELLTVSPVIVLSKKLIQFAP
IGLVNMLNTGGAIQSMEFDNHIDVVKIGVRGCGEMKVFASEKPVSCKLDG
VVVKFDYEDKMLRVQVPWPSASKLSMVEFLF

As shown in Table 3, the identities between RFSs and SIPs are about 40%. The identities between RFSs and STSs range from about 40% to about 50%. On the other hand the identities among RFSs are 60% or higher. The identities among STSs are also 60% or higher. That is, the identities among RFSs or the identities among STSs are higher than the identities between RFSs and SIP or the identities between RFSs and STSs. Thus, RFSs, SIP or STSs can be distinguished based on the results of analysis using BLAST program.

3) Attached Table 4 shows the identities obtained using another BLAST program for the amino acid sequences of RFSs, SIP and STSs shown in Table 1. All possible pair-wised amino acid sequence comparison were made by the "Blast 2 Sequences" program from NCBI (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html). Sequence identities were calculated using default parameters, program; blastp, matrix; BLOSUM62, open gap penalty; 11, extension gap penalty; 1, gap x_dropoff; 50, expect; 10.0, and word size; 3. The amino acid sequences of the RFSs, SIP and STSs used to calculate sequence identities are identical to those used as the "query" to obtain identities shown in Table 3. Results were essentially the same with former two types of comparison.

4) In conclusion, raffinose synthases (RFSs), seed imbibition protein (SIP) and stachyose synthases (STSs) were clearly distinguished from one another based on comparison of their amino acid sequences.

7. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonments, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the above-identified application or any patent issued thereon.

This 6th day of September, 2005

Akitsu NAGASAWA

Table 1

Code	Protein*	Organism	Accession**	Reference	Author/Assignee
Sc-03	RFS	*Beta vulgaris*	E37133	09/301,766	Sumitomo Chemical
Sc-05	RFS	*Brassica juncea*	E36417	09/301,766	Sumitomo Chemical
Sc-02	RFS	*Vicia faba*	E24423	08/992,914	Sumitomo Chemical
Sc-04	RFS	*Glycine max*	E24424	08/992,914	Sumitomo Chemical
Aj-05	RFS	*Cucumis sativus*	AF073744	Family GH36***	Ohsumi et al.
PsRFS	RFS	*Pisum sativum*	AJ426475	Family GH36	Peterbauer et al.
HvSIP	SIP	*Hordeum vulgare*	M77475	Family GH36	Heck et al.
PsSTS-1	STS	*Pisum sativum*	AJ311087	Family GH36	Peterbauer et al.
PsSTS-2	STS	*Pisum sativum*	AJ512932	Family GH36	Peterbauer et al.
VaSTS	STS	*Vigna angularis*	Y19024	Family GH36	Peterbauer et al.
AmSTS	STS	*Alonsoa meridionalis*	AJ487030	Family GH36	Voitsekhovskaja
SsSTS	STS	*Stachys affinis*	AJ344091	Family GH36	Pesch and Schmitz

*Protein: RFS, Raffinose synthase; SIP, Seed Imbibition Protein; STS, Stachyose synthase.

**Accession: GenBank Accession Number.

***Family GH36: glycoside hydrolase family 36 (see Carbohydrate-Active Enzymes (CAZy) database: http://afmb.cnrs-mrs.fr/CAZY/GH_36.html)

Table 2

Code	Sc-02	Sc-03	Sc-04	Sc-05	PsRFS	Aj-05	HvSIP	AmSTS	PsSTS-1	PsSTS-2	SaSTS	VaSTS
amino acids	799	783	611	777	798	784	757	868	853	847	863	857
Sc-02		62	54	62	89	64	39	43	43	42	42	43
Sc-03			50	63	63	66	39	42	43	42	42	43
Sc-04				50	54	50	29	35	35	35	34	36
Sc-05					62	70	39	44	44	43	44	45
PsRFS						64	38	42	43	43	42	43
Aj-05							38	42	44	43	43	43
HvSIP								33	34	34	34	34
AmSTS									64	64	83	64
PsSTS-1										98	65	75
PsSTS-2											65	74
SaSTS												65
VaSTS												

Table 3

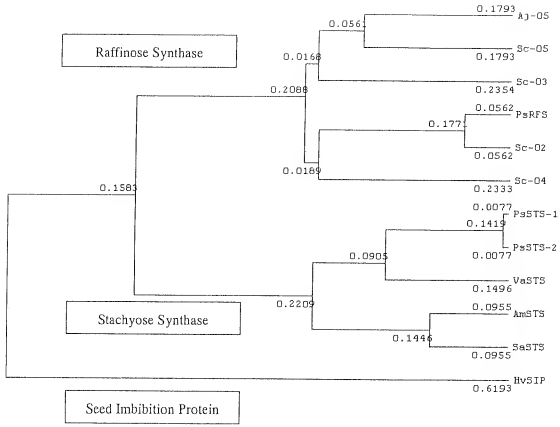
Code	Sc-02	Sc-03	Sc-04	Sc-05	PsRFS	Aj-05	HvSIP	AmSTS	PsSTS-1	PsSTS-2	SaSTS	VaSTS
amino acids	799	783	781	777	798	784	757	868	853	847	863	857
Sc-02		61	75	61	88	63	41	52	50	50	49	50
Sc-03			65	63	62	66	40	49	39	39	49	49
Sc-04				67	75	65	38	39	50	43	51	52
Sc-05					60	70	40	51	51	51	52	50
PsRFS						65	40	51	49	49	51	50
Aj-05							39	49	50	50	50	50
HvSIP								38	38	38	38	38
AmSTS									63	64	81	62
PsSTS-1										96	63	72
PsSTS-2											64	72
SaSTS												64
VaSTS												

Table 4

Code	Sc-02	Sc-03	Sc-04	Sc-05	PsRFS	Aj-05	HvSIP	AmSTS	PsSTS-1	PsSTS-2	SaSTS	VaSTS
amino acids	799	783	781	777	798	784	757	868	853	847	863	857
Sc-02		61	75	61	88	63	41	43	42	42	41	42
Sc-03			64	63	62	66	40	41	42	42	41	42
Sc-04				67	75	65	38	42	43	44	42	46
Sc-05					60	70	40	45	45	45	45	45
PsRFS						65	40	42	42	42	42	42
Aj-05							39	41	43	43	42	43
HvSIP								33	35	35	33	34
AmSTS									62	62	78	62
PsSTS-1										96	63	73
PsSTS-2											64	73
SaSTS												64
VaSTS												

Fig. 1

[GENETYX : Evolutionary tree]
Date : 2004.2.4
Method: UPGMA



 Applicant's Copy 09/301,766

The Cellulose Synthase Superfamily[1]

Todd A. Richmond* and Chris R. Somerville

Carnegie Institution of Washington, Department of Plant Biology, 260 Panama Street, Stanford, California 94305 (T.A.R., C.R.S.); and Department of Biological Sciences, Stanford University, Stanford, California 94305 (C.R.S.)

The availability of a nearly complete genome sequence for Arabidopsis has created many novel opportunities to identify, by computational methods, the genes that encode enzymes, which have been difficult to characterize by conventional means. We have used this approach to identify a large family of genes of unknown function that show sequence similarity to cellulose synthase. Our working hypothesis is that these genes encode enzymes that catalyze the synthesis of non-cellulosic polysaccharides (Cutler and Somerville, 1997).

A recent breakthrough in research concerning the biogenesis of plant cell walls was the identification, by genomic methods, of genes encoding cellulose synthase in cotton fibers (Pear et al., 1996; Delmer, 1999). The cotton cellulose synthase genes, now termed *CesA1* and *CesA2*, were identified in a collection of expressed sequence tag (EST) sequences on the basis of weak sequence similarity to genes for cellulose synthase from bacteria. In addition, the genes were expressed at high levels in cotton fibers at the onset of secondary wall synthesis and a purified fragment of one of the corresponding proteins was shown to bind UDP-Glc, the proposed substrate for cellulose biosynthesis. The conclusion that the cotton *CesA* genes are cellulose synthases is supported by results obtained with two cellulose-deficient Arabidopsis mutants, *rsw1* (Arioli et al., 1998) and *irx3* (Turner and Somerville, 1997; Taylor et al., 1999). The genes corresponding to the *RSW1* and *IRX3* loci exhibit a high degree of sequence similarity to the cotton *CesA* genes and are considered orthologs. Ten full-length *CesA* genes have been sequenced from Arabidopsis, and there is a genome survey sequence that may indicate one additional family member (Fig. 1).

It is not known at this time whether other polypeptides are also required for cellulose synthase activity (i.e. the CesA polypeptides may be a component of a multisubunit enzyme complex). Until this matter is resolved we consider it expedient to simply refer to the CesA family members as cellulose synthase. The observation that *IXR3* (AtCesA7), which is required for secondary wall cellulose synthesis, is in a different branch of the *CesA* tree than *RSW1* (AtCesA1), which is required for primary wall synthesis (Fig. 1), may indicate that there is sequence divergence between the enzymes involved in primary and secondary wall synthesis.

Reiterative database searches using the Arabidopsis Rsw1 (AtCesA1) and the cotton CesA polypeptide sequences as the initial query sequences revealed a large superfamily of at least 41 *CesA*-like genes in Arabidopsis. Based on predicted protein sequences, we have grouped these genes into seven clearly distinguishable families (Fig. 1): the *CesA* family, which includes *RSW1* and *IRX3* (*AtCesA7*), and six families of structurally related genes of unknown function designated as the "cellulose synthase-like" genes (*CslA*, *CslB*, *CslC*, *CslD*, *CslE*, and *CslG*). The nomenclature for these families is still under discussion (http://mbclserver.rutgers.edu/CPGN/CelluloseWeb/CesA.proposal.html), so the *Csl* designation for these genes should be considered temporary and may be revised as the enzymatic function of the members of each family is determined.

All of the members of the cellulose synthase superfamily appear to be integral membrane proteins, with three to six transmembrane domains in the carboxy terminal region of the protein and one or two transmembrane domains in the amino terminal region. It is thought that the CesA proteins are located in the plasma membrane (Delmer, 1999). If the Csl proteins participate in the synthesis of non-cellulosic polysaccharides, they would be expected to be located in the Golgi apparatus. Preliminary analysis of CslB, CslG, and CslE fusions to green fluorescent protein appear to localize to the Golgi (T. Richmond and C. Somerville, unpublished data). Also, immunolocalization studies with an antibody to the CslA protein indicates that this family is localized to the cytoplasm (i.e. the Golgi apparatus) rather than the plasma membrane (N. Sprenger and C. Somerville, unpublished data).

Intron-exon organization is conserved among the *CesA*, *CslB*, *CslG*, and *CslE* gene families, but not the *CslA*, *CslC*, or *CslD* families (Fig. 2). However, the C-terminus of a subset of the *CslD* genes is congruent with this organization as well. The *CslD* gene family is the most similar of the *Csl* gene families to the *CesA* family (approximately 45% identical at the amino acid level). The gene structure for this family is unusual in that the seven genes for which complete genomic sequence information is available have four

[1] This work was supported in part by the U.S. Department of Energy (grant no. DOE-FG02-00ER20133).

* Corresponding author; e-mail todd@andrew2.stanford.edu; fax 650-325-6857.

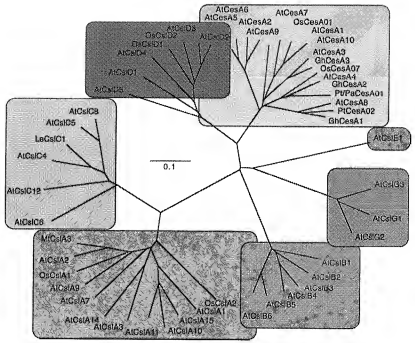


Figure 1. Unrooted, bootstrapped tree of the *CesA* superfamily. ClustalX (version 1.8) was used to create an alignment of the full-length, publicly available protein sequences that was then bootstrapped (n = 5,000 trials) to create the final tree. Subfamilies are boxed. At, Arabidopsis; Gh, cotton; Le, tomato; Mt, *Medicago truncatula*; Os, rice; Pt, *Populus tremuloides*; Pt/Pa, *Populus tremula* × *Populus alba*.

different patterns of intron-exon organization. Based on recent thinking about the evolution of intron/exon structure (de Souza et al., 1998), the small number of introns in this family, and their divergent nature, would seem to suggest that this gene family is the oldest in the cellulose synthase superfamily and may predate the *CesA* family.

All members of the *CesA* family contain a putative LIM-like Zn-binding domain/RING finger domain in the N-terminal region, which is similar to several putative plant Leu zipper transcription factors (Kawagoe and Delmer, 1997a, 1997b; Arioli et al., 1998). LIM domains are known to mediate protein-to-protein interactions (Bach, 2000), whereas RING finger domains are thought to play a role in ubiquitin-mediated proteolysis (Freemont, 2000). These domains may play a role in mediating CesA function via protein partners or targeted degradation. All of the *Csl* proteins lack this amino terminus extension, including the *CslD* family, which contains proteins similar in size to the CesAs.

Although the various CesA and Csl proteins vary in their degree of sequence similarity to one another (Table I), they share several features that have been proposed to be indicative of processive glycosyltransferases (Saxena et al., 1995). All of the *CesA* and *Csl* gene products contain a D,D,D,QxxRW motif (Fig. 2), which has been proposed to define the nucleotide sugar-binding domain and the catalytic site of these enzymes. Based on this motif, the proposed topology of these proteins (discussed above), and sequence-based classification, the various members of the Arabidopsis cellulose synthase superfamily appear to belong to family 2 of the inverting nucleotide-

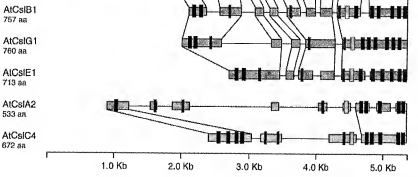


Figure 2. Comparison of the gene structure of representative genes of the Arabidopsis *CesA* superfamily. Colored boxes represent exons and the lines connecting them denote introns. Thick vertical black bars indicate predicted transmembrane domains as predicted by HMMTOP (http://www.enzim.hu/hmmtop/). Thin blue bars represent conserved Asp residues, and the thicker gray bar represents the QxxRW domain. Thin lines connecting different genes indicate conserved intron-exon junctions.

Table I. *Identity/similarity matrix for selected members of the CesA superfamily*

Identity	Similarity						
	AtCesA1	AtCslD3	AtCslB1	AtCslG1	AtCslE1	AtCslA2	AtCslC4
AtCesA1	–	48.1	31.3	29.3	30.7	13.1	14.3
AtCslD3	37.1	–	28	27.7	28.3	12.2	14
AtCslB1	22.1	18.9	–	37.4	41.1	17.4	18.6
AtCslG1	21.2	18.4	25.4	–	48.7	16.3	17.3
AtCslE1	21.4	18.9	30.1	34.4	–	16.6	18.2
AtCslA2	7.1	6.3	9.1	8.4	8.7	–	44.8
AtCslC4	8.2	6.7	9.3	8.7	9.1	31.9	–

diphospho-sugar glycosyltransferases (Campbell et al., 1997) that synthesize repeating β-glycosyl unit structures. To date, this family includes over 500 putative members, including cellulose synthase, chitin synthase, hyaluronan synthase, β-1,3-glucan synthase, and a number of uncharacterized genes from many organisms (Campbell et al., 1997; http://afmb.cnrs-mrs.fr/~pedro/CAZY/gtf_2.html). The function of the various *Csl* families is not known, but speculation is that they are responsible for producing some of the other polysaccharides found in plant cell walls and in secretions such as root cap or stylar mucilage (Cutler and Somerville, 1997). Although the D,D,D,QxxRW motif is thought to be indicative of processive β-glycosyltransferases, there is no comparative sequence data available on processive α-glycosyltransferases. Therefore we cannot rule out the possibility that some of these enzymes produce polysaccharides with α-linkages, such as rhamnogalacturonan I or rhamnogalacturonan II. It is possible that linkage specificity is determined by subtle features in the active site of the proteins (Stasinopoulos et al., 1999) and that members of the Arabidopsis cellulose synthase superfamily make polysaccharides with both β- and α-linkages.

DISCUSSION

With six families of *Csl* genes and six major noncellulosic polysaccharides in Arabidopsis (i.e. callose, xyloglucan, glucuronoarabinoxylan, homogalacturonan, rhamnogalacturonan I, and rhamnogalacturonan II), it is tempting to speculate that each family is responsible for the biosynthesis of one of the principal polysaccharides of the cell wall. Although we consider it possible that the gene superfamily described here encodes enzymes that catalyze the synthesis of different polymers, there is at present no evidence for this other than the observation that sequence divergence is frequently associated with functional divergence. It is also possible that there are additional functional divisions within the gene families that are not evident from our analysis. Recent results concerning the relationship between enzyme structure and function, such as experiments showing that as few as four amino acid changes can alter the catalytic outcome of an enzymatic reaction from desaturation to hydroxylation (Broun et al., 1998), emphasize the need for caution in using sequence similarity to infer function based on sequence.

The amount of plant genome sequence and EST information in the public sequence databases is expanding rapidly. At present there are more than 900,000 plant ESTs and genome survey sequences in GenBank, most of which are from 35 species. In the first 8 months of the year 2000, more than 516,000 new ESTs and genome survey sequences from 16 plant species were deposited. Thus except for species such as Arabidopsis, which will soon be completely sequenced, any attempt at a comprehensive compilation of CesA-related sequence information represents a continuing challenge. To facilitate research on these genes, we have established a website (http://cellwall.stanford.edu) that summarizes the ever-increasing number of cellulose synthase and cellulose synthase-like genes. At present, there are more than 1,250 *CesA* and *Csl* sequences, from 29 different plant species in GenBank. Although the most extensive information available is for Arabidopsis where there are more than 330 partial or complete gene sequences, there is also a significant amount of information available for several other species, especially rice, maize, soybean, and tomato. A crude estimate of the relative abundance of

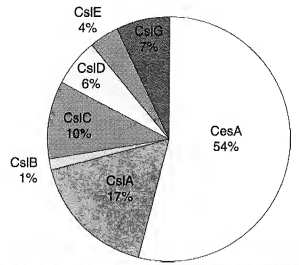


Figure 3. Relative abundance of EST sequences for members of the *CesA* and *Csl* families in GenBank.

mRNA for the various family members can be calculated from the frequency with which each gene family is represented by EST sequences in the public databases (Fig. 3).

Polysaccharides found in other plant species, but not in Arabidopsis (Zablackis et al., 1995), such as mixed linkage xylans, mannans, or arabinans, may be synthesized by genes that are not represented by orthologs in Arabidopsis. A number of gene sequences from plants in GenBank show limited similarity (<50% identity) to the members of the various *Csl* families in Arabidopsis. This and other issues will undoubtedly become more transparent when the function of the *Csl* genes in Arabidopsis is known from direct experimental evidence. Our laboratory, along with others, is examining the patterns of gene expression and protein localization of the Arabidopsis *Csl* genes, and attempting to characterize their enzymatic function using reverse genetics. We are confident that in the next several years the function of these genes will be understood and it will then be possible to begin to unravel the challenge of understanding how cell wall composition and deposition is controlled.

Received May 25, 2000; accepted July 7, 2000.

LITERATURE CITED

Arioli T, Peng L, Betzner AS, Burn J, Wittke W, Herth W, Camilleri C, Hofte H, Plazinski J, Birch R, Cork A, Glover J, Redmond J, Williamson RE (1998) Molecular analysis of cellulose biosynthesis in *Arabidopsis*. Science **279:** 717–720

Bach I (2000) The LIM domain: regulation by association. Mech Dev **91:** 5–17

Broun P, Shanklin J, Whittle E, Somerville CR (1998) Catalytic plasticity of fatty acid modification enzymes underlying chemical diversity of plant fatty acids. Science **282:** 1315–1317

Campbell JA, Davies GJ, Bulone V, Henrissat B (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem J **326:** 929–939

Cutler S, Somerville C (1997) Cellulose synthase: cloning by in silico. Curr Biol **7:** R108–R111

Delmer DP (1999) Cellulose biosynthesis: exciting times for a difficult field of study. Annu Rev Plant Physiol Plant Mol Biol **50:** 245–276

de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W (1998) Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA **95:** 5094–5099

Freemont PS (2000) Ubiquitination: ring for destruction? Curr Biol **10:** R84–R87

Kawagoe Y, Delmer DP (1997a) Cotton CelA1 has a LIM-like Zn binding domain in the N-terminal cytoplasmic region (abstract no. 337). Plant Physiol **114:** S–85

Kawagoe Y, Delmer DP (1997b) Pathways and genes involved in cellulose biosynthesis. Genet Eng **19:** 63–87

Pear JR, Kawagoe Y, Schreckengost WE, Delmer DP, Stalker DM (1996) Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. Proc Natl Acad Sci USA **93:** 12637–12642

Saxena IM, Brown RM Jr, Fevre M, Geremia RA, Henrissat B (1995) Multidomain architecture of β-glycosyl transferases: implications for mechanism of action. J Bacteriol **177:** 1419–1424

Stasinopoulos SJ, Fisher PR, Stone BA, Stanisich VA (1999) Detection of two loci involved in (1→3)-β-glucan (curdlan) biosynthesis by *Agrobacterium* sp. ATCC31749, and comparative sequence analysis of the putative curdlan synthase gene. Glycobiology **9:** 31–41

Taylor NG, Scheible WR, Cutler S, Somerville CR, Turner SR (1999) The *irregular xylem 3* locus of Arabidopsis encodes a cellulose synthase gene required for secondary cell wall synthesis. Plant Cell **11:** 769–780

Turner SR, Somerville CR (1997) Collapsed xylem phenotype of Arabidopsis identifies mutants deficient in cellulose deposition in the secondary cell wall. Plant Cell **9:** 689–701

Zablackis E, Huang J, Müller B, Darvill AG, Albersheim P (1995) Characterization of the cell wall polysaccharides of *Arabidopsis thaliana* leaves. Plant Physiol **107:** 1129–1138

ELSEVIER

Gene 190 (1997) 245–249

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

Applicants Copy 09/301,764

Identification of an acetolactate synthase small subunit gene in two eukaryotes

Ronald G. Duggleby *

Centre for Protein Structure, Function and Engineering, Department of Biochemistry, University of Queensland, Brisbane 4072, Australia

Received 30 September 1996; received in revised form 27 November 1996; accepted 28 November 1996

Abstract

Acetolactate synthase catalyses the first step in branched-chain amino acid biosynthesis. The bacterial enzyme contains two large and two small subunits but there is only limited and circumstantial evidence for a small subunit in the cukaryotic enzyme. Here this evidence is summarised and protein sequences of two putative eukaryotic small subunits, from a yeast and a red alga, are presented. © Elsevier Science B.V. All rights reserved.

Keywords: Branched-chain amino acid; Chloroplast genome; Herbicide; Yeast genome

1. Introduction

Acetolactate synthase (ALS) is an essential enzyme in plants and many microorganisms because it catalyses the first step in the biosynthesis of branched-chain amino acids. In some bacteria it also plays a catabolic role, supplying acetolactate for the butanediol fermentation.

There appear to be two distinct forms of the enzyme that correspond to these functional roles. The anabolic enzyme contains FAD (Schloss et al., 1985) and is inhibited by the branched-chain amino acids (Weinstock et al., 1992) while the catabolic enzyme, sometimes referred to as the 'pH 6 acetolactate-forming enzyme', displays neither of these properties (Störmer, 1968; Peng et al., 1992). A further property of the anabolic enzyme is that it is inhibited by a number of compounds that are used as herbicides (Schloss et al., 1988). The remainder of this article concerns the anabolic enzyme only.

Many of the bacterial ALSs have been shown to be heterotetramers composed of two types of subunit, large and small. The latter subunit was first identified (Squires et al., 1983) for *Escherichia coli* isoenzyme III (ALSIII); DNA sequencing revealed an open reading frame that appeared to have a homologue in the operon that contains the gene for *E. coli* ALSII (Lawther et al., 1981). The protein product of the small subunit gene was later identified for *E. coli* ALSI (Eoyang and Silverman, 1984) and *Salmonella typhimurium* ALSII (Schloss et al., 1985).

The role of the small subunit is not entirely clear and it may be that it is involved in more than one way. For the various *E. coli* isoforms it has been shown that this subunit affects sensitivity to branched-chain amino acids (Eoyang and Silverman, 1986; Sella et al., 1993), specific activity (Lu and Umbarger, 1987), stability (Sella et al., 1993) and the kinetic properties (Weinstock et al., 1992).

Putative small subunit genes have been identified for a number of other bacterial species. This identification has been based mainly, and in most cases solely, on the presence of an open reading frame 3′ to the large subunit gene. In contrast, the presence of a small ALS subunit has never been demonstrated unequivocally in eukaryotes. Certainly no open reading frame nearby the large subunit gene has been identified but this is not surprising since operons are not a feature of eukaryotic genomes. However, there is some evidence that a small subunit may exist.

First, purified wheat ALS contains a low molecular weight component (Southan and Copeland, 1996) that could be a small subunit; on the other hand, it could be simply an impurity. Purified barley ALS has been reported to contain no small subunit (Durner and Böger, 1988) on the basis of SDS-PAGE. However, it is conceiv-

* Corresponding author. Tel.: +61 7 33654615; Fax: +61 7 33654699; e-mail: duggleby@biosci.uq.edu.au

Abbreviations: ALS, acetolactate synthase.

PII S0378-1119(97)00002-4

```
Bfl     ----------------MANSDVTRHILSVLVQDVDGIISRVSGMFTRRAFNLVSLVSAKT   44
Ccr     MTANVQPAPASAYDLSPKDQAEQSATFALLVDNEPGVLHRVVGLFAARGYNIESLTVAET   60
Cgl     ----------------MANSDVTRHILSVLVQDVDGIISRVSGMFTRRAFNLVSLVSAKT   44
EcoH    ----------------------MRRILSVLLENESGALSRVIGLFSQRGYNIESLTVAPT   38
EcoM    ----------------------MQHQVNVSARFNPETLERVLRVVRHRGFHVCSMNMAAA   38
EcoN    ----------------MQNTTHDNVILELTVRNHPGVMTDVCGLFARRAFNVEGILCLPI   44
Mav     -------------------MSPQTHTLSVLVEAKPGVLARVAALFSRRGFNIESLAVGAT   41
Sav     ---------------------MSKHTLSVLVENKPGVLARITALFSRRGFNIDSLAVGVT   39
Sty     ----------------------MRRILSVLLENESGALSRVIGLFSQRGYNIESLTVAPT   38
                                    .        .  .  .   * ...  .

Bfl     E-THGINRITVVVD-ADELNIEQITKQLNKLIPVLKVVRLDEETT-IARAIMLVKVSADS  101
Ccr     DRKAHTSRITVVTR-GTRHVLDQIEAQLNKVVNVRRVHDVTRDPNGVERELALVKVRGSG  119
Cgl     E-THGINRITVVVD-ADELNIEQITKQLNKLIPVLKVVRLDEETT-IARAIMLVKVSADS  101
EcoH    D-DPTLSRMTIQTV-GDEKVLEQIEKQLHKLVDVLRVSELGQGAH-VEREIMLVKIQASG   95
EcoM    S-DAQNINIELTVA--SPRSVDLLFSQLNKLVDVAHVAICQSTTT--SQQIRA-------   86
EcoN    Q-DSDKSHIWLLVN--DDQRLEQMISQIDKLEDVVKVQRNQSDPTMFNKIAVFFQ-----   96
Mav     E-QKDMSRMTIVVS-AEETPLEQITKQLNKLINVIKIVELEDGNS-VSRELALIKVRADA   98
Sav     E-HPDISRITIVVNVIEALPLEQVTKQLNKLVNVLKIVELEPSAGRAGGELVLVKVRADN   98
Sty     D-DPTLSRMTIQTV-GDEKVLEQIEKQLHKLVDVLRVSELGQGAH-VEREIMLVKMEASG   95
                . .         .. .  *. *.  * ..

Bfl     TNRPQIVDAANIFRARVVDVAPDSVVIESTGTPGKLRALLDVMEPFG-IRELIQSGQIAL  160
Ccr     VDRLEALRIAEIFRAKPVDTTLESFVFEISGAPSKIDKFLDLMRPLG-LVELSRTGVLSI  178
Cgl     TNRPQIVDAANIFRARVVDVAPDSVVIESTGTPGKLRALLDVMEPFG-IRELIQSGQIAL  160
EcoH    YGRDEVKRNTEIFRGQIIDVTPSLYTVQLAGTSGKLSAFLASIRDVAKIVEVARSGVVGL  155
EcoM    ------------------------------------------------------------   86
EcoN    ------------------------------------------------------------   96
Mav     GTRSQVIEAVNLFRAKVIDVSPEALTIEATGDRGKIEALLRVLEPSV-SVRSS-NREWCR  156
Sav     ETRSQIVEIVQLFRAKTVDVSPEAVTIEATGGSDKLEAMLKMLEPFR-HQGARQSGTIAI  157
Sty     YGREEVKRNTEIFRGQIIDVTPTLYTVQLAGTSDKLDAFLASLRDVAKIVEVARSGVVGL  155

Bfl     NRGPKTMAPAKI-------                                          172
Ccr     ERGFEGM------------                                          185
Cgl     NRGPKTMAPAKI-------                                          172
EcoH    SRGDKIMR-----------                                          163
EcoM    -------------------                                           86
EcoN    -------------------                                           96
Mav     CPGPRGIGTAK--------                                          167
Sav     GRGARSITDRSLRPLDRSA                                          176
Sty     SRGDKIMR-----------                                          163
```

Fig. 1. Alignment of selected ALS small subunit protein sequences. Sequences were obtained from GenBank and aligned using the ClustalW (Thompson et al., 1994) program. An asterisk indicates a totally conserved residue, while a full stop denotes a position where there are conservative substitutions. Abbreviations used are: Bfl, *Brevibacterium flavum* MJ233; Ccr, *Caulobacter crescentus*; Cgl, *Corynebacterium glutamicum*; EcoH, *E. coli ilvH* (ALSIII); EcoM, *E. coli ilvM* (ALSII); EcoN, *E. coli ilvN* (ALSI); Mav, *Mycobacterium avium*; Sav, *Streptomyces avermitilis*; Sty, *S. typhimurium*.

able that it could be lost during multistep purification; in this context it is relevant that the various *E. coli* isoforms have differing affinities for their respective small subunits and that, for ALSIII, the small subunit is readily lost (Sella et al., 1993). In addition, even when a small subunit is present, it is not easily observed by SDS-PAGE (De Rossi et al., 1995) because it migrates as a rapidly moving, diffuse band that stains only weakly with Coomassie blue.

Second, we have confirmed (Chang and Duggleby, unpublished) that expression of the *Arabidopsis thaliana* ALS-encoding gene in *E. coli* results in an enzyme that, unlike the enzyme from the plant itself, is insensitive to inhibition by branched-chain amino acids (Singh et al., 1992). The suggested explanation (Singh et al., 1992) is that the expressed enzyme lacks a small subunit, although no evidence was adduced to support this proposal. A number of other explanations of this observation are possible, such as different post-translational processing, including proteolysis, between prokaryotes and eukaryotes. The plant enzyme is located in the chloroplast and contains an amino-terminal sequence that is believed to be a chloroplast transit peptide (Mazur et al., 1987). Although the enzyme expressed in *E. coli* is processed to a similar size as the native enzyme (Singh et al., 1992), it is not known whether cleavage of the transit peptide is at the same site as in the plant. Expression of the yeast enzyme in *E. coli* also results in an enzyme that is kinetically distinguishable from the native enzyme (Poulsen and Stougaard, 1989); this difference has also been ascribed to the lack of the appropriate small subunit.

Third, over-expression of the *A. thaliana* ALS-encoding gene in tobacco (Odell et al., 1990) or oilseed rape (Ouellet et al., 1994) gives greatly elevated amounts of the corresponding mRNA, but much smaller increases in ALS activity. This lack of correlation could be interpreted to indicate that some other component, such as a small subunit, is limiting.

Although none of these lines of evidence for an ALS

```
Bfl  ------------------------------------------------------------  0
Ppu  ------------------------------------------------------------  0
Sce  MLRSLLQSGHRRVVASSCATMVRCSSSSTSALAVKQMHRHATRPPLPTLDTPSWNANSAV  60

Bfl  ---------MANSDVTRHILSVLVQDVDGIISRVSGMPTRRAFNLVSLVSAKTETHGINR  51
Ppu  ------------MKHTLSVLVQDEAGVLSRISGLPARRGFNIASLAVGPAEQIGVS-  44
Sce  SSIIYETPAPSRQPRKQHVLNCLVQNEPGVLSRVSGTLAARGFNIDSLVVCNTEVKDLSR  120
                  * *  *** + *+++* **  + *+**+ ** +   *    +

Bfl  ITVVVDADELNIEQITKQLNKLIPVLKVVRLDEETTIARAIMLVKVSAD----------  100
Ppu  ITMVVQGDNRTIEQLTKQLYKLVNILNVQDVTNIPSVERELMLIKIQVN----------  93
Sce  MTIVLQGQDGVVEQARRQIEDLVPVYAVLDYTNSEIIKRELVMARISLIGTEYFEDLLLH  180
      * * ++      **   *   *+    * + ++       *++     +

Bfl  ----------STNRPQIVD-------------------------AANIFRARVVDVA  122
Ppu  ----------SQNRIEALEVK-----------------------IFRANVVDIA  114
Sce  HHTSTNAGAADSQELVAEIREKQFHPANLPASEVLRLKHEHLNDITNLTNNPGGRVVDIS  240
                #+          +                      *    ***+

Bfl  PDSVVIESTGTPGKLRALLDVMEPFGIRELIQSGQIALNRGPK--TMAPAKI--------  172
Ppu  EDLIIVEVTGDPGKIVAIEQLLTKFGIIEIARTGKISLVRTSKINTYLKDKVVAYNA---  171
Sce  ETSCIVELSAKPTRISAFLKLVEPFGVLECARSGMMALPRTPLKTSTEEAADEDEKISEM  300
     +   +++*    * + *   +   **  * ++ *   * *+

Bfl  ---------                                                    172
Ppu  ---------                                                    171
Sce  VDISQLPPG                                                    309
```

Fig. 2. Alignment of a bacterial and two putative eukaryotic ALS small subunit protein sequences. The *B. flavum* MJ233 (Bfl), *P. purpurea* (Ppu) and *S. cerevisiae* (Sce) sequences were obtained from GenBank and aligned using the ClustalW (Thompson et al., 1994) program. An asterisk indicates a totally conserved residue, while a plus denotes a position where the two eukaryotic sequences are identical. The underlined segment represents the most highly conserved region.

small subunit in eukaryotes is alone convincing, taken together they suggest that further work is merited. Recent advances in genome sequencing have provided an opportunity to look for ALS small subunit genes. Here the presence of such genes in two eukaryotic species is reported.

2. Results and discussion

There is limited similarity between the protein sequences of known ALS small subunits as illustrated in Fig. 1, which shows an alignment of several such sequences. In all, only four residues are totally conserved and a further 17 positions show conservative substitutions. A consensus sequence was derived from this alignment and compared to the individual sequences. The *B. flavum* and *C. glutamicum* sequences are most similar to the consensus and the former was used as a representative small subunit sequence. Various databases in GenBank were searched for conceptual translations into protein sequences that are similar to this *B. flavum* small subunit, using the BLAST program. In addition to known bacterial ALS small subunit genes, this search identified two well-matched eukaryotic genes: one (with a probability of arising by chance of 8.2×10^{-42}) from the chloroplast genome of the red alga *Porphyra purpurea* (Reith and Munholland, 1995) and the other ($P = 2.7 \times 10^{-25}$) from chromosome III of the yeast *Saccharomyces cerevisiae* (Oliver et al., 1992). The next best match ($P = 1.0$) corresponded to a fragment of a mouse lipoxygenase gene (Chen et al., 1994), mistranslated in a reverse reading frame. The *S. cerevisiae* and *P. purpurea* sequences are shown in Fig. 2, aligned with that of the ALS small subunit of *B. flavum*. These results clearly indicate that *S. cerevisiae* and *P. purpurea* contain a gene that could encode an ALS small subunit. A total of 38 residues are identical in all three sequences and, when only the two eukaryotes are compared, there are 66 identities. The most highly conserved region is a 26 residue sequence near the amino terminus, with the motif LVQXXXGϕϕSRϕSGXXXXRXFNϕXSL, where X is any amino acid while ϕ is one of V, L or I.

The *S. cerevisiae* sequence is substantially longer than the others; compared to that of *P. purpurea*, there are an additional 50 residues in the middle of the protein, a short extension at the carboxyl-terminus, and a 75 residue extension at the amino-terminus. It is suggested that the role of the latter is to act as a mitochondrial transit peptide, since it is known that *S. cerevisiae* ALS is a mitochondrial enzyme (Ryan and Kohlhaw, 1974). The transit peptides of other mitochondrial proteins frequently contain an arginine residue at position −2 relative to the cleavage site (von Heijne et al., 1989) and it is noted that this putative *S. cerevisiae* ALS small subunit contains an arginine at position 75, close to where homology with the *P. purpurea* sequence begins. Thus, it is proposed that the cleavage site is immediately after K76. It has also been noted (von Heijne et al., 1989) that mitochondrial transit peptides are enriched in A, L, R and S, but deficient in D and E. Similar characteristics are observed here, most notably for S

(which constitutes 18.4% of the first 76 residues but only 7.3% of the remaining 233 residues) and E (1.3% versus 9.4%).

Unlike ALS large subunits from plants (Mazur et al., 1987), the proposed *P. purpurea* ALS small subunit does not contain a chloroplast transit sequence. However, this is not necessary as the gene is located in the chloroplast genome. Thus it is suggested that in this plant, the large subunit is synthesised in the cytoplasm and transported to the chloroplast where it associates with the chloroplast-encoded small subunit. This arrangement is very similar to the situation often observed for ribulose 1,6-*bis*phosphate carboxylase, except that in that case it is the larger of the two subunits that is encoded by the chloroplast genome (Spreitzer, 1993).

Finding what appears to be an ALS small subunit gene in two eukaryotes as diverse as a yeast and a red alga suggests that small subunit genes will exist in other plants and fungi. However, the location of this gene, as well as that for the large subunit, may be variable. For example, it has been shown that in another red alga, *P. umbilicus*, an ALS large subunit is encoded by a chloroplast gene (Reith and Munholland, 1993). Further, the location of the *P. purpurea* ALS small subunit gene in the chloroplast may be unusual. We have searched for this gene in the complete chloroplast genomes of five other plants: *Nicotiana tabacum* (Shinozaki et al., 1986), *Oryza sativa* (Hiratsuka et al., 1989), *Pinus thunbergii* (Tsudzuki et al., 1992), *Marchantia polymorpha* (Ohyama et al., 1986) and *Odontella sinensis* (Kowallik et al., 1995). A total of 608 open reading frames were examined but the best match with the motif mentioned previously contained only 8 of the 17 conserved residues and bore no overall similarity to ALS small subunits; in contrast, the three sequences in Fig. 2 match in all 17 positions.

Because ALS is the target for several herbicides (Schloss et al., 1988), there has been considerable interest in transforming crop plants with herbicide-resistant forms of the enzyme (Odell et al., 1990; Ouellet et al., 1994). The success of this procedure is likely to be limited if a small subunit is an essential component of the plant enzyme. Thus, the work reported here may have significant practical implications. At present, there is no evidence that ALS small subunit genes exist in any eukaryotic species apart from *S. cerevisiae* and *P. purpurea*, or that even in these species the genes are actually expressed. Indeed, it is possible that these two genes serve an entirely different function that is unrelated to ALS activity. Ultimately the function of any DNA sequence, whose identity is based solely on homology, can only be proven by experiments designed to evaluate that function. In the case of these putative eukaryotic ALS small subunit genes, their function might be demonstrated by gene disruption or by co-expression with the large subunit genes. Current studies in this laboratory are examining these possibilities.

3. Note added in proof

Recent examination of GenBank expressed sequence tags has identified three sequences (two from *A. thaliana* and one from rice) that may represent higher plant ALS small subunits. The last of these gives a very good match to the *P. purpurea* sequence; over residues 83–154 there are 46 identical, and 10 similar, amino acids. This EST is apparently encoded in the nucleus, as it is not present in the rice chloroplast genome.

References

Chen, X.S., Kurre, U., Jenkins, N.A., Copeland, N.G. and Funk, C.D. (1994) J. Biol. Chem. 269, 13979–13987.

De Rossi, E., Leva, R., Gusberti, L., Manachini, P.L. and Riccardi, G. (1995) Gene 166, 127–132.

Durner, J. and Böger, P. (1988) Z. Naturforsch. 43c, 850–856.

Eoyang, L. and Silverman, P.M. (1984) J. Bacteriol. 157, 184–189.

Eoyang, L. and Silverman, P.M. (1986) J. Bacteriol. 166, 901–904.

Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., Kondo, C., Honji, Y., Sun, C.R., Meng, B.Y. et al. (1989) Mol. Gen. Genet. 217, 185–194.

Kowallik, K.V., Stoebe, B., Schaffran, I., Kroth-Pancic, P. and Freier, U. (1995) Plant Mol. Biol. Rep. 13, 336–342.

Lawther, R.P., Calhoun, D.H., Adams, C.W., Hauser, C.A., Gray, J. and Hatfield, G.W. (1981) Proc. Natl. Acad. Sci. USA 78, 922–925.

Lu, M.F. and Umbarger, H.E. (1987) J. Bacteriol. 169, 600–604.

Mazur, B.J., Chui, C.F. and Smith, J.K. (1987) Plant Physiol. 85, 110–117.

Odell, J.T., Caimi, P.G., Yadav, N.S. and Mauvais, C.J. (1990) Plant Physiol. 94, 1647–1654.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H. and Ozeki, H. (1986) Nature 322, 572–574.

Oliver, S.G., van der Aart, Q.J.M., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P.G., Benit, P. et al. (1992) Nature 357, 38–46.

Ouellet, T., Mourad, G., Brown, D., King, J. and Miki, B. (1994) Plant Sci. 102, 91–94.

Peng, H.L., Wang, P.Y., Wu, C.M., Hwang, D.C. and Chang, H.Y. (1992) Gene 117, 125–130.

Poulsen, C. and Stougaard, P. (1989) Eur. J. Biochem. 185, 433–439.

Reith, M.E. and Munholland, (1993) J. Curr. Genet. 23, 59–65.

Reith, M.E. and Munholland, J. (1995) Plant Mol. Biol. Rep. 13, 333–335.

Ryan, E.D. and Kohlhaw, G.B. (1974) J. Bacteriol. 120, 631–637.

Schloss, J.V., Ciskanik, L.M. and Van Dyk, D.E. (1988) Nature 331, 360–362.

Schloss, J.V., Van Dyk, D.E., Vasta, J.F. and Kutny, R.M. (1985) Biochemistry 24, 4952–4959.

Sella, C., Weinstock, O., Barak, Z. and Chipman, D.M. (1993) J. Bacteriol. 175, 5339–5343.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K. et al. (1986) EMBO J. 5, 2043–2049.

Singh, B., Szarnosi, I., Hand, J.M. and Misra, R. (1992) Plant Physiol. 99, 812–816.

Southan, M.D. and Copeland, L. (1996) Physiol. Plant. 96, 171–179.
Spreitzer, R.J. (1993) Annu. Rev. Plant. Physiol. Mol. Biol. 44, 411–434.
Squires, C.H., DeFelice, M., Deveraux, J. and Calvo, J.M. (1983) Nucleic Acids Res. 11, 5299–5313.
Störmer, F.C. (1968) J. Biol. Chem. 243, 3740–3741.
Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
Tsudzuki, J., Nakashima, K., Tsudzuki, T., Hiratsuka, J., Shibata, M., Wakasugi, T. and Sugiura, M. (1992) Mol. Gen. Genet. 232, 206–214.
von Heijne, G., Steppuhn, J. and Hermann, R.G. (1989) Eur. J. Biochem. 180, 535–545.
Weinstock, O., Sella, C., Chipman, D.M. and Barak, Z. (1992) J. Bacteriol. 174, 5560–5566.

Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

JAMES U. BOWIE,* JOHN F. REIDHAAR-OLSON, WENDELL A. LIM, ROBERT T. SAUER

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

THE GENOME IS MANIFEST LARGELY IN THE SET OF PROteins that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is fundamental to an understanding of biological processes. Because an amino acid sequence contains all of the information necessary to determine the structure of a protein (*1*), it should be possible to predict structure from sequence, and subsequently to infer detailed aspects of function from the structure. However, both problems are extremely complex, and it seems unlikely that either will be solved in an exact manner in the near future. It may be possible to obtain approximate solutions by using experimental data to simplify the problem. In this article, we describe how an analysis of allowed amino acid substitutions in proteins can be used to reduce the complexity of sequences and reveal important aspects of structure and function.

Methods for Studying Tolerance to Sequence Variation

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. This method has been extremely powerful for proteins such as the globins or cytochromes, for which sequences from many different species are known (*2–7*). The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (*3, 8–11*). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions (*2–4, 11*). For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in *lac* repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent (*11*). At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a 10^4-fold reduction in activity (*12*). A similar loss of activity occurs in λ repressor when a DNA binding residue is changed from Asn to Asp (*13*). To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity (*10, 14–16*). Hence, many of the residues that are conserved in sets of related sequences play structural roles.

Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved (*6*). Similar results have been seen for a number of protein families (*2, 4, 5, 7, 17, 18*). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in λ repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the protein (*9*). These substitutions were identified by a functional

The authors are in the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

*Present address: Department of Chemistry and Biochemistry and the Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90024.

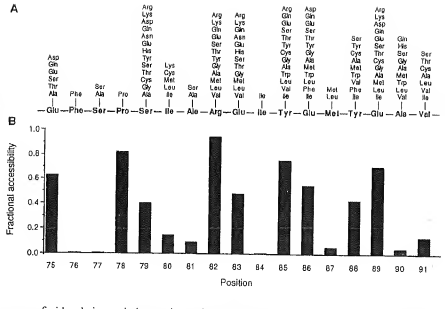


Fig. 1. (**A**) Amino acid substitutions allowed in a short region of λ repressor. The wild-type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutating one to three codons at a time by using a cassette method and applying a functional selection (*9*). (**B**) The fractional solvent accessibility (*42*) of the wild-type side chain in the protein dimer (*43*) relative to the same atoms in an Ala-X-Ala model tripeptide.

selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried by the residues that are solvent inaccessible.

Constraints on Core Sequences

Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability (*19*). For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the NH_2-terminal domain of λ repressor (*20*). The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited (*21*). Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied (*22*).

The cores of most proteins are quite closely packed (*23*), but some volume changes are acceptable. In λ repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core position in the appropriate sequence contexts. Large volume changes at individual buried sites have also been observed in phylogenetic studies, where it has been noted that the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion (*5, 7, 17*). Rather, local volume changes are accommodated by conformational changes in nearby side chains and by a variety of backbone movements.

The Informational Importance of the Core

With occasional exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the core is composed of side chains that can assume only a limited number of conformations (*24*), efficient packing must be maintained without steric clashes. How important are hydrophobicity, volume, and steric complementarity in determining whether a given sequence can form an acceptable core? Each factor is essential in a physical sense, as a stable core is probably unable to tolerate unsatisfied hydrogen bonding groups, large holes, or steric overlaps (*25*). However, in an informational sense, these factors are not equivalent. For example, in experiments in which three core residues of λ repressor were mutated simultaneously, volume was a relatively unimportant informational constraint because three-quarters of all possible combinations of the 20 naturally occurring amino acids had volumes within the range tolerated in the core, and yet most of these sequences were unacceptable (*20*). In contrast, of the sequences that contained only

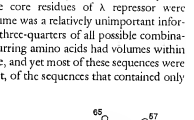


Fig. 2. Amino acid substitutions allowed in the core of λ repressor. The wild-type side chains are shown pictorially in the approximate orientation seen in the crystal structure (*43*). The lists of allowed substitutions at each position are shown below the wild-type side chains. These substitutions were identified by randomly mutating one to four residues at a time by using a cassette method and applying a functional selection (*20*). Not all substitutions are allowed in every sequence background.

40	36	47	65	51	57	18
Ala	Ala	Ala	Ala	Cys	Cys	Ala
Cys	Cys	Cys	Ser	Val	Pro	Cys
Val	Thr	Thr	Cys	Ile	Val	Val
Ile	Val	Val	Thr	Leu	Ile	Ile
Leu	Ile	Ile	Val	Met	Leu	Leu
Met	Leu	Leu	Ile	Phe	Met	Met
Phe	Met	Met	Leu		Phe	
	Phe	Phe	Met			
			Phe			

the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (*4*), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in λ repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of active sequences, but it is not possible to decide whether a side chain is functionally or structurally important just because it is invariant or conserved. To make this distinction requires an independent assay of protein folding. The ability of a mutant protein to maintain a stably folded structure can often be measured by biophysical techniques, by susceptibility to intracellular proteolysis (*26*), or by binding to antibodies specific for the native structure (*27, 28*). In the latter cases, it is possible to screen proteins in mutated clones for the ability to fold even if these proteins are inactive. Sets of sequences that allow formation of a stable structure can then be compared to the sets that allow both folding and function, with the active site or binding residues being those that are variable in the set of stable proteins but invariant in the set of functional proteins. The DNA-binding residues of Arc repressor were identified by this method (*8*). The receptor-binding residues of human growth hormone were also identified by comparing the stabilities and activities of a set of mutant sequences (*28*). However, in this case, the mutants were generated as hybrid sequences between growth hormone and related hormones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (*29, 30*). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have

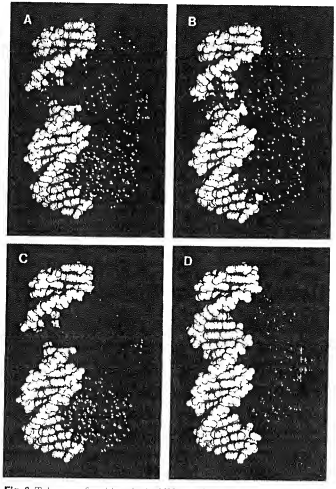


Fig. 3. Tolerance of positions in the NH_2-terminal domain of λ repressor to hydrophilic side chains. The complex (*43*) of the repressor dimer (blue) and operator DNA (white) is shown. In (**A**), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (**B**) without the remaining protein atoms. In (**C**), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (**D**) without the remaining protein atoms. About three-fourths of the 92 side chains in the NH_2-terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (*9, 14, 30, 27, 44*).

been used to combine such information into more appropriately weighted sequence searches and alignments (*31*). These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 (*29*). Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure (*32*).

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected (*33, 34*). These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences (*6*). This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots for the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain proteins (Fig. 4B) (*35*). The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (*34*). The amphipathic character of the three α-helical regions in the Antennapedia protein (*36*) is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of Arc repressor, a small DNA-binding protein, was recently predicted by a similar method (*8*) and confirmed by nuclear magnetic resonance studies (*37*).

The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of different structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion for their structural relatedness (*38*). In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure (*4, 29, 38, 39*). Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure de novo. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple sequences composed only of H (hydrophobic) and P (polar) groups on two-dimensional lattices (*40*). An example of such a representation is shown in Fig. 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice, and two residues cannot occupy the same space. Free energies of particular conformations are evaluated with a single term, an attraction of H groups. By considering chains of ten residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence or conformation space could be explored. The significant results were as follows: (i) not all sequences can fold into a "native" structure and only a few sequences form a unique native structure; (ii) the probability that a sequence will adopt a unique native structure increases with chain length; and (iii) the native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. Although the gap between these two-dimensional simulations and three-dimensional structures is large, the use of simple rules and sequence representations yields results similar to those expected for real proteins. Three-dimensional lattice methods are also beginning to be developed and evaluated (*41*).

Summary

There is more information in a set of related sequences than in a single sequence. A number of practical applications arise from an analysis of the tolerance of residue positions to change. First, such information permits the evaluation of a residue's importance to the function and stability of a protein. This ability to identify the essential elements of a protein sequence may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy

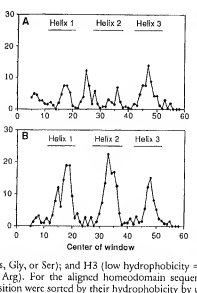


Fig. 4. Helical hydrophobic moments calculated by using (**A**) the Antennapedia homeodomain sequence or (**B**) a set of 39 aligned homeodomain sequences (*35*). The bars indicate the extent of the helical regions identified in nuclear magnetic resonance studies of the Antennapedia homeodomain (*36*). To determine hydrophobic moments, residues were assigned to one of three groups: H1 (high hydrophobicity = Trp, Ile, Phe, Leu, Met, Val, or Cys); H2 (medium hydrophobicity = Tyr, Pro, Ala, Thr, His, Gly, or Ser); and H3 (low hydrophobicity = Gln, Asn, Glu, Asp, Lys, or Arg). For the aligned homeodomain sequences, the residues at each position were sorted by their hydrophobicity by using the scale of Fauchere and Pliska (*45*). Arg and Lys were not counted unless no other residue was found at the position, because they contain long aliphatic side chains and can thereby substitute for nonpolar residues at some buried sites. To account for possible sequence errors and rare exceptions, the most hydrophilic residue allowed at each position was discarded unless it was observed twice. The second most hydrophilic residue was then chosen to represent the hydrophobicity of each position. An eight-residue window was used and the vectors projected radially every 100°. The vector magnitudes were assigned a value of 1, 0, or −1 for positions where the hydrophobicity group was H1, H2, or H3, respectively.

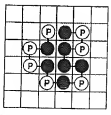


Fig. 5. A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (*40*), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further expcriments and more sophisticated energy calculations.

REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, F. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965)
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. J. H. Miller *et al.*, *J. Mol. Biol.* **131**, 191 (1979).
12. S. Sprang *et al.*, *Science* **237**, 905 (1987); C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* **26**, 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (*5*) have argued that a protein core composed solely of hydrogen bonded residues would also be invisible on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989); W. S. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Breyer and R. T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330 (1989).
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown *et al.*, *J. Mol. Biol.* **42**, 65 (1969); J. Greer, *ibid.* **153**, 1027 (1981); J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia *et al.*, *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim, *J. Mol. Biol.* **88**, 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982), D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988)
36. G. Otting *et al.*, *EMBO J.* **7**, 4305 (1988).
37. J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. R. Vershon, R. T. Sauer, D. J. Patel, *ibid.*, p. 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.*, in preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987), D. G. Covell and R. L. Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369 (1983).

46. We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.U.B.) and Howard Hughes Medical Institute (W.A.L.).

MOLECULAR AND CELLULAR BIOLOGY, Mar. 1988, p. 1247–1252
0270-7306/88/031247-06$02.00/0

Vol. 8, No. 3

Transforming Growth Factor α: Mutation of Aspartic Acid 47 and Leucine 48 Results in Different Biological Activities

ELIANE LAZAR,† SHINICHI WATANABE,* STEPHEN DALTON, AND MICHAEL B. SPORN

Laboratory of Chemoprevention, National Cancer Institute, Bethesda, Maryland 20892

Received 22 July 1987/Accepted 30 November 1987

To study the relationship between the primary structure of transforming growth factor α (TGF-α) and some of its functional properties (competition with epidermal growth factor (EGF) for binding to the EGF receptor and induction of anchorage-independent growth), we introduced single amino acid mutations into the sequence for the fully processed, 50-amino-acid human TGF-α. The wild-type and mutant proteins were expressed in a vector by using a yeast α mating pheromone promoter. Mutations of two amino acids that are conserved in the family of the EGF-like peptides and are located in the carboxy-terminal part of TGF-α resulted in different biological effects. When aspartic acid 47 was mutated to alanine or asparagine, biological activity was retained; in contrast, substitutions of this residue with serine or glutamic acid generated mutants with reduced binding and colony-forming capacities. When leucine 48 was mutated to alanine, a complete loss of binding and colony-forming abilities resulted; mutation of leucine 48 to isoleucine or methionine resulted in very low activities. Our data suggest that these two adjacent conserved amino acids in positions 47 and 48 play different roles in defining the structure and/or biological activity of TGF-α and that the carboxy terminus of TGF-α is involved in interactions with cellular TGF-α receptors. The side chain of leucine 48 appears to be crucial either indirectly in determining the biologically active conformation of TGF-α or directly in the molecular recognition of TGF-α by its receptor.

Transforming growth factor α (TGF-α) is a polypeptide of 50 amino acids. First isolated from a retrovirus-transformed mouse cell line (9), it has subsequently been found in human tumor cells (10, 29), in the early rat embryo (18), and recently in cell cultures from the pituitary gland (23). TGF-α appears to be closely related to epidermal growth factor (EGF) structurally and functionally (19, 20). The two peptides apparently bind to the same receptor, and both induce anchorage-independent growth of certain nontransformed cells, such as NRK cells, in the presence of TGF-β (1).

Comparison of amino acid sequences reveals about 35% homology among the EGF-like peptides (rat [27], mouse [25], and human [13] EGFs and rat [19] and human [12] TGF-αs). Some viral peptides (Shope fibroma growth factor [6], vaccinia growth factor [2], and myxoma growth factor [30]) also share homologies with the EGF-like peptides.

If TGF-α is involved in transformation, a TGF-α antagonist could be an important therapeutic tool in the treatment of certain types of malignancies. An understanding of the conformational and dynamic properties of the TGF-α molecule is basic to the design of an antagonist. A hypothetical antagonist would bind to the same receptor as TGF-α, but would not induce the series of proliferative and transforming events induced by TGF-α. To obtain such a molecule it is necessary to dissociate interactions responsible for binding from those involved in signal transduction. We decided to approach the problem by way of site-directed mutagenesis of a human sequence of TGF-α. In this report we describe our first series of mutations, which were carried out at residues Asp-47 and Leu-48, in the carboxy-terminal part of TGF-α; these two amino acids are highly conserved in the EGF-like family of peptides. We show that these two adjacent residues play different roles in the structure and/or function of TGF-α.

MATERIALS AND METHODS

Cells. Normal rat kidney (NRK) cells were grown in Dulbecco modified Eagle medium containing 10% (vol/vol) calf serum.

TGF-α gene. The sequence of the 50-amino-acid human TGF-α was originally derived from a human TGF-α precursor cDNA (12). The coding sequence is preceded by an ATG methionine codon and followed by a TAA stop codon and is flanked by *Eco*RI restriction sites. This *Eco*RI fragment combines the 59-base-pair *Eco*RI-*Nco*I fragment from plasmid pTE5 (12) with the 111-base-pair *Nco*I-*Eco*RI fragment from plasmid pyTE2 (11). The resulting *Eco*RI fragment was inserted in M13mp18 for site-directed mutagenesis.

Synthesis and purification of oligonucleotides and oligonucleotide-directed mutagenesis. The synthesis and purification of 20- to 27-nucleotide oligonucleotides were carried out as described previously (31). The one or two nucleotides responsible for the mutation were located in the middle of the oligonucleotide. Mutagenesis was performed by published procedures (21, 33). The sequences of the mutant clones were verified by the method of Sanger et al. (25).

Yeast shuttle vector. The vector YEp70αT contains a yeast α-factor pheromone promoter and prepro sequence for the expression of TGF-α (15). The mutant TGF-α coding sequence was inserted in the *Eco*RI site of plasmid YEp70αT and expressed in the form of a fusion protein consisting of 92 amino acids from the prepro sequence of the yeast α factor attached to the amino terminus of TGF-α (28). The yeast cleaves the precursor and secretes TGF-α with 8 amino acids fused to it (4 are encoded by the prepro sequence of α-factor, and the other 4 are encoded by the DNA sequence added to insert of the TGF-α gene). The last of these residues is a methionine, which allows the cleavage of the secreted fusion

* Corresponding author.
† Present address: Unité d'Oncologie Moleculaire, IRSC, 94802 Villejuif Cedex, France.

protein by cyanogen bromide (CNBr) and the release of a mature TGF-α (50 amino acids) (see Results).

Yeast strain and transformation. The yeast *Saccharomyces cerevisiae* 20B-12 (*MATα trp1 pep4-3*) (17) was obtained from the Yeast Genetics Stock Center, Berkeley, Calif. *S. cerevisiae* 20B-12 was grown in YEPD medium (1% yeast extract [Difco Laboratories], 2% Bacto-Peptone [Difco], 2% glucose). When the culture reached an optical density at 660 nm of 1, spheroplasts were prepared (14) for transformation. For each transformation we used 10 to 15 μg of purified plasmid DNA.

Partial purification of TGF-α mutants. At 3 days after transformation, five individual colonies of transformants were grown to saturation in YEPD medium. The amount of protein in the yeast medium was measured by the method of Bradford (3), and the amount of mutant TGF-α secreted in the yeast medium was determined by radioimmunoassay. The clones which secrete the highest amount of mutant TGF-α were used to grow a 1-liter culture in YNB-CAA medium (0.67% yeast nitrogen base, 20 g of glucose per liter, 10 g of Casamino Acids [Difco] per liter). After the culture reached saturation (optical density at 660 nm of 10 to 12) (48 h in an air shaker at 30°C), the yeast conditioned medium was dialyzed extensively against 1 M acetic acid in 3,000-molecular-weight cutoff dialysis tubing. Usually 250 ml of dialyzed culture was lyophilized, suspended in 10 ml of 70% formic acid, and treated with CNBr (molar excess of 500) for 20 h at room temperature. The CNBr was subsequently evaporated, and the samples were lyophilized. CNBr-treated samples were suspended in 1 ml of 1 M acetic acid, loaded on a Bio-gel P30 column (30 by 1.5 cm [Bio-Rad Laboratories]), and eluted with 1 M acetic acid. Fractions of 1 ml were collected. Aliquots were lyophilized, suspended in binding buffer (minimum essential medium containing 1 mg of bovine serum albumin per ml and 25 mM HEPES [*N*-2-hydroxyethylpiperazine-*N'*-2-ethanesulfonic acid; pH 7.4]), neutralized if necessary to pH 7.4, and tested in EGF-binding competition and soft-agar assays, as well in radioimmunoassay.

Radioimmunoassays. The amounts of TGF-α secreted in the yeast medium were determined by radioimmunoassay with the immunoglobulin G fraction of a polyclonal antibody, 34D, raised against recombinant human TGF-α (4), in 0.1 M Tris (pH 7.5)–0.15 M NaCl–2.5 mg of bovine serum albumin per ml. The amounts of partially purified TGF-α present in the P30 column fractions were measured by using the Biotope RIA kit with polyclonal antibody against human TGF-α (a gift from W. Hargreaves, Biotope), under denaturing conditions, as recommended by the supplier.

EGF binding competition assay and soft agar assay. Both EGF-binding competition and soft-agar assays have been described previously (1).

RESULTS

Rationale for mutations in the carboxyl terminus of TGF-α. Figure 1 shows the amino acid sequence of TGF-α in which the residues that are conserved among all the EGF-like peptides described thus far (EGF, TGF-α, and EGF-like viral proteins) are enclosed in bold circles. Among the 11 conserved amino acids, there are 6 Cys and 2 Gly residues, which presumably play essential roles in determining the overall conformation of the molecule. We concentrated on the two conserved amino acids in the carboxyl terminus, Asp-47 and Leu-48. The Asp in position 47 is conserved among the EGFs and TGF-α (human or murine), but not among the EGF-like viral proteins (vaccinia growth factor, Shope fibroma growth factor, or myxoma growth factor), whereas Leu 48 is conserved among all the EGF-like peptides so far described. In both mouse and human EGF, the two corresponding residues (Asp-46 and Leu-47) are located near the surface of the protein (8, 22, 22a). We designed a series of mutations in these two positions.

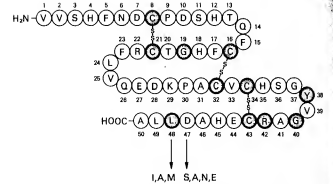


FIG. 1. Mutations in the carboxy terminus of human TGF-α. The amino acids conserved in all the family of EGF-like growth factors (human and murine EGFs and TGFs, as well as the gene products of the vaccinia virus [vaccinia growth factor], the Shope fibroma virus [Shope fibroma growth factor], and the myxoma virus [myxoma growth factor]) are enclosed in bold circles. The mutations of amino acids at positions 47 and 48 are indicated. Symbols: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.

Asp-47 has been mutated to Glu, Asn, Ser, and Ala. Glu was chosen because it has the same charge as and a larger size than Asp; Asn has a similar side-chain structure, but is uncharged; Ser is smaller but still polar; Ala is smaller and nonpolar.

Leu 48 has been mutated to Ile and Met, which are both large, nonpolar residues like Leu, and to Ala, which is nonpolar but smaller. We introduced the chosen mutations by site-directed mutagenesis of the cloned human TGF-α gene, using synthetic oligonucleotides.

Construction of the yeast α mating pheromone–human TGF-α plasmid. The TGF-α expression vector pyTE1 (Fig. 2) was constructed by using plasmid YEp70αT (15) which contains the 2μm origin of replication and yeast *TRP1* gene for its replication and selective maintenance, respectively. YEp70αT also contains the yeast α-factor promoter, the α-factor prepro sequence coding for 89 amino acids, and the sequence for 3 amino acids resulting from the introduction of *Xba*I and *Eco*RI sites. The human mature TGF-α sequence (12) is contained in a 170-base-pair *Eco*RI fragment which includes an ATG (Met) codon preceding the sequence of TGF-α and a TAA (stop) codon followed by 8 nucleotides. This TGF-α sequence was inserted in the unique *Eco*RI site of YEp70αT. Clones with the proper orientation were selected, and DNA was isolated for yeast transformation.

Measurement of TGF-α secreted by *S. cerevisiae*. The amount of total proteins secreted into the yeast culture was 10 ± 1 μg/ml for wild-type as well as mutant TGF-α as determined by the method of Bradford (3). Before further purification was attempted, we wanted to determine whether the mutated TGF-α proteins were being secreted by the yeast. The low pH of the yeast medium, as well as the acidic proteins secreted in the yeast culture, precluded biological assay of secreted mutants. Therefore, immunological meth-

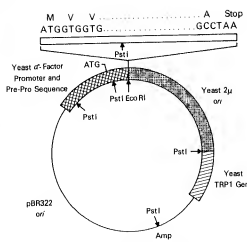


FIG. 2. Structure of the *S. cerevisiae* 8.2-kilobase shuttle vector pyTE1. The secretion of the TGF-α gene is under the transcriptional control of the yeast α-factor promoter and prepro sequence ([illegible]). The yeast 2μm origin of replication ([illegible]) and the selective yeast *TRP1* gene ([illegible]) are indicated. The TGF-α gene, preceded by an initiation (ATG) codon and followed by a stop (TAA) codon, is inserted in the *Eco*RI site. Details are given in Materials and Methods and in Results.

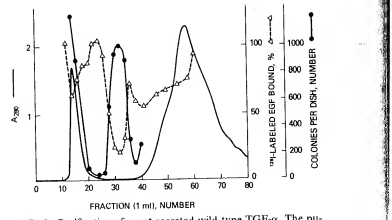


FIG. 3. Purification of yeast-secreted wild-type TGF-α. The purification procedure is described in Materials and Methods and in Results. Aliquots of every other fraction of the Bio-Gel P30 column were tested for their abilities to compete with ^{125}I-EGF for binding to the EGF receptor (△) and to induce colony formation (>62 μm) on NRK cells in soft agar in the presence of TGF-β (1 ng/ml) (●). The A_{280} profile of the proteins was determined (——).

ods were used. Wild-type and mutant TGF-α's were secreted at a level of 100 to 200 ng/ml and 10 to 500 ng/ml, respectively (as determined by radioimmunoassay with polyclonal antibody 34D). We thus estimate that the percentage of TGF-α secreted in the yeast culture is at least 1% of the total protein secreted. We cannot yet assess whether the variations in the levels of secretion of different mutant TGF-α proteins are real or whether one single-amino-acid substitution drastically affects the recognition by the antibody. The latter hypothesis is the more likely, since the use of another polyclonal antibody (Biotope) under denaturing conditions enabled us to detect certain TGF-α mutants (such as [Ala 47]-TGF-α, in which the amino acid in position 47 of human TGF-α is mutated to an alanine) that were poorly detected by 34D, under nondenaturing as well as denaturing conditions. After the amount of TGF-α mutant proteins was estimated, the medium was extensively dialyzed against 1 M acetic acid and lyophilized as described in Materials and Methods.

Partial purification of yeast-secreted TGF-α. Although the yeast shuttle vector was constructed in such a way as to secrete TGF-α with 8 amino acids fused to the N terminus, it was often observed that a significant fraction of the secreted TGF-α was in a higher-molecular-weight fragment corresponding to the size expected from an uncleaved (unprocessed) 92-amino-acid fusion protein. Since a Met had been introduced at the N terminus of TGF-α and since TGF-α contains no Met in its sequence, CNBr treatment could be used to cleave either of these 8- or 92-amino-acid N-terminal peptides and release the complete 50-amino-acid TGF-α. Indeed, CNBr treatment of yeast-secreted proteins resulted in the conversion of high-molecular-weight TGF-α into the 6,000-molecular-weight species, as revealed by Western immunoblot (data not shown).

CNBr-cleaved samples (see Materials and Methods) were purified on a Bio-Gel P30 column. Figure 3 shows the elution profile of the proteins, as well as the results of a radioreceptor assay and a soft-agar assay performed on aliquots of the column fractions. The A_{280} profile shows two major peaks of eluted proteins, one corresponding to the void volume and the other one to proteins of molecular weight <3,000. Aliquots of the column fractions were tested for their ability to compete with ^{125}I-EGF for binding to the receptor. The fractions that were the most active in this assay were located between the two major protein peaks, in an area where relatively few proteins eluted. Although some activity was found in the first protein peak (void volume), this was considerably reduced on treatment with stronger CNBr (data not shown).

Aliquots of each fraction were also tested for their ability to induce anchorage-independent growth of NRK cells in soft agar in the presence of TGF-β (1 ng/ml). The receptor binding and colony-forming activity superimposed almost exactly (Fig. 3). Analysis by polyacrylamide gel electrophoresis with silver staining, as well as by Western blot, of the column fractions shows that our purification procedure (CNBr cleavage followed by P30 sizing column) eliminates high-molecular-weight proteins (data not shown). Since pure TGF-α migrates in a broad band on sodium dodecyl sulfate-polyacrylamide gel electrophoresis (32), this technique cannot be used for proper assessment of the degree of separation of TGF-α from low-molecular-weight contaminating proteins. Nevertheless, within our detection levels the amounts of TGF-α present in the column fractions (detected by radioimmunoassay using the antibody from Biotope) correlated with the amounts observed on sodium dodecyl sulfate-polyacrylamide gel electrophoresis (data not shown).

Comparison of binding and colony-forming activity of TGF-α partially purified from yeast media. It was important to show that wild-type TGF-α secreted from *S. cerevisiae* had the expected biological properties and that its activity in soft-agar and radioreceptor assays was equivalent. For these assays, the amount of EGF-competing activity present in the most active fraction of the P30 column of wild-type TGF-α was measured in terms of EGF equivalents. The dilution curve had a slope that was parallel to that of the EGF standard. This value was also used to measure the colony-forming activity of the partially purified wild-type TGF-α (with EGF as a standard in the assay). The colony-forming activity of the partially purified wild-type TGF-α corre-

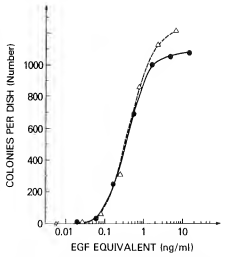


FIG. 4. Correlation between the activities in the binding and colony-forming assay for the partially purified wild-type TGF-α secreted by *S. cerevisiae*. The activity in the radioreceptor assay of the peak fraction from the P30 column was determined in EGF equivalent concentration. The value obtained was used for the soft-agar assay. Colonies of >62 μm (△) and the EGF standard (●) are shown.

sponded exactly to that of EGF (Fig. 4). Thus, we have partially purified a wild-type 50-amino-acid TGF-α showing the expected binding and colony-forming activities, which provides a reference substance for mutant TGF-αs that might show a dissociation of binding and colony-forming abilities.

Biological and biochemical activities of the partially purified TGF-α mutant proteins. Mutated TGF-αs were expressed by using the yeast system and partially purified on Bio-Gel P30 columns as described in Materials and Methods. Mutant TGF-αs were usually obtained from two different clones of yeast transformants. The CNBr-cleaved samples were purified through different Bio-Gel P30 columns for each mutant protein to avoid any possible contamination from one peptide to another. The purification profiles observed with the mutant TGF-αs were similar to those obtained for the wild-type TGF-α. Aliquots of the P30 column fractions were tested in radioreceptor and soft-agar assays. For all mutant proteins, the highest activity in both assays was always found in the same fraction of the Bio-Gel P30 column effluent (peak fraction). Extensive purification of a series of mutant proteins for screening purposes is not practical. Therefore, we needed a quantitation system that would allow us to compare mutant proteins with each other. Thus, the amount of TGF-α present in the peak fraction was estimated by radioimmunoassay with an antiserum to native TGF-α (obtained from W. Hargreaves), under denaturing conditions, as described in Materials and Methods. All values given in Table 1 were obtained from the peak fraction.

The controls done with the wild-type TGF-α showed (Fig. 4; Table 1) that binding and transforming activity were equivalent. The yeast vector without a TGF-α insert did not secrete any EGF-like proteins, as determined by both radioreceptor and soft-agar assay.

Two types of results were obtained upon assay of mutant proteins having different amino acid substitutions at Asp-47. In both [Ala-47]–TGF-α and [Asn-47]–TGF-α, binding ability was retained. Soft-agar and radioreceptor activities correlated for [Asn-47]–TGF-α; there was a lower value for

TABLE 1. Biological and biochemical activities of mutant TGF-α proteins secreted by *S. cerevisiae* and partially purified

Insert in the yeast expression vector	EGF equivalence (ng/ml) in:		Amt of TGF-α (ng/ml) in radioimmunoassay
	Radioreceptor assay	Soft-agar assay	
Wild-type TGF-α	700	700	2,000
	400	300	ND[a]
None	0	0	0
[Ala-47]–TGF-α	100	44	220
	66	48	ND
[Asn-47]–TGF-α	80	72	180
	75	72	525
[Glu-47]–TGF-α	3	3	42
[Ser-47]–TGF-α	10	4	60
[Ala-48]–TGF-α	0	0	16
	0	0	220
[Ile-48]–TGF-α	4	12	470
	2	7	490
[Met-48]–TGF-α	2	8	453
	0.5	2	420

[a] ND, Not determined.

colony-forming activity than for EGF-binding competition for [Ala-47]–TGF-α. [Ser-47]–TGF-α and [Glu-47]–TGF-α appeared to have lower activities in both assays than either wild-type TGF-α or [Ala-47]–TGF-α and [Asn-47]–TGF-α. These results indicate that neither the carboxyl charge nor the polarity of Asp-47 is essential for biological activity.

The effects of mutation of Leu-48, one of the 11 amino acids perfectly conserved among all the EGFs, TGF-αs, and viral EGF-like proteins, are dramatic. [Ala-48]–TGF-α totally lacked binding and colony-forming activity. [Ile-48]–TGF-α and [Met-48]–TGF-α had very little biological activity compared with wild-type TGF-α. Another substitution, [Met-48]–TGF-α, resulted in a truncated mutant lacking the last 2 amino acids and having a substitution of Leu to homoserine at position 48 following treatment with CNBr. Alternatively, if [Met-48]–TGF-α was not treated with CNBr, fusion proteins of TGF-α (mutated to Met in position 48) with 8 or 92 amino acids attached at the N terminus were obtained. Very low activities in binding and soft-agar assays were found for these mutants, whether or not they were cleaved with CNBr. Experiments on EGF and TGF-α have shown that an N-terminal extension does not markedly modify EGF-binding activity (12, 26). Therefore, the loss of activity obtained with [Met-48]–TGF-α that has not been CNBr treated was probably due to the mutation itself and not to the N-terminally extended fusion protein. We do not know whether the loss of activity observed with the TGF-α shortened to 48 amino acids and having a substitution of Leu-48 to homoserine is due only to the mutation or also to the lack of the last 2 amino acids.

The data obtained by radioimmunoassay on the partially purified wild-type and mutant TGF-α show that the amount of TGF-α detected was always higher than the amount determined by measurement of biological activity. This may be due to the presence in the fraction of a certain percentage of incorrectly folded TGF-α that might be recognized in a

radioimmunoassay under denaturing conditions but would not be biologically active. None of the mutant proteins seemed to be present in amounts equivalent to those observed for wild-type TGF-α in the partially purified fractions (whether radioimmunoassay, radioreceptor, or soft-agar assay was used for quantitation). It is not clear whether consistently less TGF-α was produced by the mutant constructs than by the wild type or whether the secreted mutant proteins were simply less well recognized by the antibody. Because of these uncertainties, the biological activities of the different mutant proteins cannot be accurately related to a known amount of mutant TGF-α protein. Even though radioimmunoassay should be used with caution for a quantitative evaluation of mutant TGF-α proteins, a positive reaction demonstrates that immunoreactive TGF-α was present in the P30 peak fraction for each mutant. Therefore, the fact that one of the mutant proteins ([Ala-48]–TGF-α) is biologically inactive can be attributed to the mutation itself, and not to the lack of production of the mutant protein by the yeast or its loss through purification. However, if the mutant proteins are in fact as immunoreactive as the wild type, then [Ala-47]–TGF-α and [Asn-47]–TGF-α are as active as wild-type TGF-α and [Glu-47]–TGF-α and [Ser-47]–TGF-α are less active; in contrast, [Ile-48]–TGF-α and [Met-48]–TGF-α are almost inactive. The differences between mutation of Asp-47 and Leu-48 would then be even more striking.

DISCUSSION

TGF-α shows sequence homologies with EGF, and both growth factors share the same cellular receptors (20). Even though EGF was discovered 25 years ago (7) and its properties have been extensively studied over the years (5), the binding site of EGF to its receptor has still not been determined, and the relationship between structure and function of EGF/TGF-α is still to be discovered. Particularly, we do not know whether binding to the receptor and signal transduction occur through one or more domains of the molecule or through which amino acids. We approached the question by performing site-directed mutagenesis of TGF-α and focused our attention on two adjacent amino acids, Asp-47 and Leu-48, located in the carboxy terminus and highly conserved in the EGF-like family of peptides. Unexpectedly, these two amino acids showed very different sensitivities to mutation and particularly to a substitution to Ala: [Ala-47]–TGF-α retained binding and colony-forming activities, whereas [Ala-48]–TGF-α completely lost both activities. These data show that Asp-47 and Leu-48 play very different roles in defining the structure and/or the activity of TGF-α. The other mutations performed on Asp-47 were substitutions to Asn, Ser, and Glu. [Asn-47]–TGF-α, like [Ala-47]–TGF-α, was active in binding and induction of colony formation, but [Ser-47]–TGF-α and [Glu-47]–TGF-α showed weaker growth factor activities. These results indicate that neither the carboxyl charge nor the polarity of Asp-47 is essential for biological activity. Interestingly, two of the EGF-like viral proteins, myxoma growth factor and Shope fibroma growth factor (6, 30), have Asn instead of Asp in position 47; we have shown that [Asn-47]–TGF-α retains biological activity.

Substitution of Leu-48 to Met and Ile led to mutant proteins with very low activities, whereas substitution to Ala led to complete loss of activity. We did not expect that a mutation of Leu to Ile (which have similar sizes and polarities) would cause such a strong effect. Thus, Leu-48, which is conserved perfectly among all the EGF-like peptides, seems to be essential, through its exact geometry, for the biological activity of TGF-α.

The mutant proteins tested so far, when active, showed parallel behaviors in binding and colony formation. Some mutant proteins lost all activities, and we assume that the binding capacity has been lost. We have not been able to dissociate the binding and colony-forming abilities by using any of the present series of mutant proteins, and it is necessary to screen more of them in search of an antagonist of TGF-α.

Results relating to the biological activity of EGF show that derivatives of mouse EGF and human EGF (EGF 1-47) lacking the carboxy-terminal 6 amino acids as a result of enzymatic digestion are less potent than the intact molecule in mitogenic stimulation of fibroblasts, but retain full biological activity in in vivo assays (inhibition of gastric acid secretion) (16). On the other hand, naturally occurring truncated forms of rat EGF, which lack the carboxy-terminal 5 amino acids (rEGF 2-48) are as potent as mouse EGF (mEGF 1-53) in receptor-binding and mitogenic assays (27). We do not know whether the discrepancies observed are due to the origin of the molecule (artificial or natural) or to the type of bioassay used. In any event, all of these EGF-related molecules, which are shorter than mouse or human EGF, still retain Leu-47. We have shown that in TGF-α, the corresponding residue, Leu-48, is critical for the biological activity.

Recent data on the three-dimensional structure of mouse EGF obtained by nuclear magnetic resonance show that even though Asp-46 and Leu-47 (Asp-47 and Leu-48 in TGF-α) are both solvent accessible (8, 22, 22a), their side chains point in opposite directions in the beta-sheet structure. Therefore, the role of these adjacent amino acids in the structure and, consequently, the function of EGF might be very different. Our data show that the amino acids Asp-47 and Leu-48 of TGF-α are not equally important for the biological activity of TGF-α, despite their conservation among the EGF-like peptides. From the dramatic loss in biological activity which is characteristic of mutation of Leu-48, we also suggest that this residue is involved in binding to the cellular receptors either by direct interaction with the receptor or by providing the proper conformation to the molecule.

ACKNOWLEDGMENTS

We thank Rik Derynck (Genentech) for providing the TGF-α gene, inserted in M13, and for his assistance throughout this project. We are indebted to Arjun Singh (Genentech) for helping us with the yeast transformation and expression. We are grateful to Tim Bringman (Genentech) and William Hargreaves (Biotope) for their generous gifts of TGF-α antibodies. We thank Linda Durham for technical assistance, Irene Dalton for manuscript preparation, and our colleagues for helpful comments and moral support.

LITERATURE CITED

1. **Anzano, M. A., A. B. Roberts, J. M. Smith, M. B. Sporn, and J. E. De Larco.** 1983. Sarcoma growth factor from conditioned medium of virally transformed cells is composed of both type α and type β transforming growth factors. Proc. Natl. Acad. Sci. USA **80**:6264–6268.
2. **Blomquist, M. C., L. T. Hunt, and W. C. Barker.** 1984. Vaccinia virus 19-kilodalton protein: relationship to several mammalian proteins, including two growth factors. Proc. Natl. Acad. Sci. USA **81**:7363–7367.
3. **Bradford, M. M.** 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. **72**:248–254.

4. **Bringman, T. S., P. B. Lindquist, and R. Derynck.** 1987. Different transforming growth factor-α species are derived from a glycosylated and palmitoylated transmembrane precursor. Cell **48**:429–440.
5. **Carpenter, G., and S. Cohen.** 1979. Epidermal growth factor. Annu. Rev. Biochem. **48**:193–216.
6. **Chang, W., C. Upton, S. Hu, A. F. Purchio, and G. McFadden.** 1987. The genome of Shope fibroma virus, a tumorigenic poxvirus, contains a growth factor gene with sequence similarity to those encoding epidermal growth factor and transforming growth factor α. Mol. Cell. Biol. **7**:535–540.
7. **Cohen, S.** 1962. Isolation of a mouse submaxillary gland protein accelerating incisor eruption and eyelid opening in the new-born animal. J. Biol. Chem. **237**:1555–1562.
8. **Cooke, R. M., A. J. Wilkinson, M. Baron, A. Pastore, M. J. Tappin, I. D. Campbell, H. Gregory, and B. Sheard.** 1987. The solution structure of human epidermal growth factor. Nature (London) **327**:339–341.
9. **De Larco, J. E., and G. J. Todaro.** 1978. Growth factors from murine sarcoma virus-transformed cells. Proc. Natl. Acad. Sci. USA **75**:4001–4005.
10. **Derynck, R., D. V. Goeddel, A. Ullrich, J. U. Gutterman, R. D. Williams, T. S. Bringman, and W. H. Berger.** 1987. Synthesis of messenger RNAs for transforming growth factors α and β and the epidermal growth factor receptor by human tumors. Cancer Res. **47**:707–712.
11. **Derynck, R., A. B. Roberts, D. H. Eaton, M. E. Winkler, and D. V. Goeddel.** 1985. Human transforming growth factor-α: precursor sequence, gene structure, and heterologous expression. Cancer Cells **3**:79–86.
12. **Derynck, R., A. B. Roberts, M. E. Winkler, E. Y. Chen, and D. V. Goeddel.** 1984. Human transforming growth factor-α: Precursor structure and expression in E. coli. Cell **38**:287–297.
13. **Gregory, H.** 1975. Isolation and structure of urogastrone and its relationship to epidermal growth factor. Nature (London) **257**:325–327.
14. **Hinnen, A., J. B. Hicks, and G. R. Fink.** 1978. Transformation of yeast. Proc. Natl. Acad. Sci. USA **75**:1929–1933.
15. **Hitzeman, R. A., C. N. Chang, M. Matteucci, L. J. Perry, W. J. Kohr, J. J. Wulf, J. R. Swartz, C. Y. Chen, and A. Singh.** 1986. Construction of expression vectors for secretion of human interferons by yeast. Methods Enzymol. **119**:424–433.
16. **Hollenberg, M. D., and H. Gregory.** 1980. Epidermal growth factor—urogastrone: biological activity and receptor binding of derivatives. Mol. Pharmacol. **17**:314–320.
17. **Jones, E.** 1976. Proteinase mutants of *Saccharomyces cerevisiae*. Genetics **85**:23–30.
18. **Lee, D. C., R. Rochford, G. J. Todaro, and L. P. Villarreal.** 1985. Developmental expression of rat transforming growth factor α mRNA. Mol. Cell. Biol. **5**:3644–3646.
19. **Marquardt, H., M. W. Hunkapiller, L. E. Hood, and G. J. Todaro.** 1984. Rat transforming growth factor type 1: structure and relation to epidermal growth factor. Science **223**:1079–1082.
20. **Massague, J.** 1983. Epidermal growth factor-like transforming growth factor. II. Interaction with epidermal growth factor receptors in human placenta membranes and A431 cells. J. Biol. Chem. **258**:13614–13620.
21. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101**:20–78.
22. **Montelione, G. T., K. Wuthrich, E. C. Nice, A. W. Burgess, and H. A. Scheraga.** 1986. Identification of two anti-parallel beta-sheet conformations in the solution structure of murine epidermal growth factor by proton magnetic resonance. Proc. Natl. Acad. Sci. USA **83**:8594–8598.

22a. **Montelione, G. J., K. Wuthrich, E. C. Nice, A. W. Burgess, and H. A. Scheraga.** 1987. Solution structure of murine epidermal growth factor: determination of the polypeptide backbone chain-fold by nuclear magnetic resonance and distance geometry. Proc. Natl. Acad. Sci. USA **84**:5226–5230.

23. **Samsoondar, J., M. S. Kobrin, and J. E. Kudlow.** 1986. α-Transforming growth factor secreted by untransformed bovine anterior pituitary cells in culture. J. Biol. Chem. **261**: 14408–14413.
24. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
25. **Savage, C. R., Jr., J. H. Hash, and S. Cohen.** 1973. Epidermal growth factor: location of disulfide bonds. J. Biol. Chem. **248**:7669–7672.
26. **Shechter, Y., J. Schlessinger, S. Jacobs, K. J. Chang, and P. Cuatrecasas.** 1978. Fluorescent labeling of hormone receptors in viable cells: preparation and properties of highly fluorescent derivatives of epidermal growth factor and insulin. Proc. Natl. Acad. Sci. USA **75**:2135–2139.
27. **Simpson, R. J., A. Smith, R. L. Moritz, M. J. O'Hare, P. S. Rudland, J. R. Morrison, C. J. Lloyd, B. Grego, A. W. Burgess, and E. C. Nice.** 1985. Rat epidermal growth factor: complete amino acid sequence. Eur. J. Biochem. **153**:629–637.
28. **Singh, A., J. M. Lugovoy, W. J. Kohr, and L. J. Perry.** 1984. Synthesis, secretion and processing of α-factor-interferon fusion proteins in yeast. Nucleic Acids Res. **12**:8927–8938.
29. **Todaro, G. J., C. Fryling, and J. E. DeLarco.** 1980. Transforming growth factors produced by certain human tumors: polypeptides that interact with epidermal growth factor receptors. Proc. Natl. Acad. Sci. USA **77**:5258–5262.
30. **Upton, C., J. L. Macen, and G. McFadden.** 1987. Mapping and sequencing of a gene from myxoma virus that is related to those encoding epidermal growth factor and transforming growth factor α. J. Virol. **61**:1271–1275.
31. **Watanabe, S., E. Lazar, and M. B. Sporn.** 1987. Transformation of normal rat kidney (NRK) cells by an infectious retrovirus carrying a synthetic rat type α transforming growth factor gene. Proc. Natl. Acad. Sci. USA **84**:1258–1262.
32. **Winkler, M. E., T. Bringman, and B. J. Marks.** 1986. The purification of fully active recombinant transforming growth factor-α produced in *Escherichia coli*. J. Biol. Chem. **261**:13838–13843.
33. **Zoller, M. J., and M. Smith.** 1983. Oligonucleotide-directed mutagenesis of DNA fragments cloned into M13 vectors. Methods Enzymol. **101**:468–500.

C12N9/02L99B C12N9/02L
C12N9/02L99

18 S. J. Mansour, J. M. Candia, J. E. Matsuura, M. C. Manning, N. G. Ahn, *Biochemistry* 35, 15529 (1996).
19 The microtubule nucleation assay was done essentially as described [T. Stearns and M. Kirschner, *Cell* 76, 623 (1994)]. Briefly, rhodamine-labeled bovine tubulin (from T. Stearns, Stanford University) and *Xenopus* sperm were added to extracts to yield final concentrations of 120 μg/ml and 200 nuclei per microliter respectively. Extracts were incubated at room temperature for 10 min. Samples were diluted in 9 vol of glutaraldehyde (0.25%), centrifuged through a 25% glycerol cushion onto coverslips, and stained with DAPI. The DAPI-stained sperm and rhodamine-labeled tubulin were examined by fluores-
20 Addition of purified *Xenopus* or rat MAPK to interphase extracts can be sufficient to produce mitotic-like microtubules under some circumstances [Y. Gotoh et al., *Nature* 349, 251 (1991)]. One difference between their experiment and ours is the way the interphase extracts were prepared. We used cycloheximide-soaked eggs, and prepared the inter-
21 N. Furuno et al., *EMBO J.* 13, 2399 (1994).
22 M. H. Verlhac et al., *Development* 122, 815 (1996).
23 Supported by a grant from the National Institutes of Health (GM46383). We thank T. Stearns for providing rhodamine-labeled tubulin and for advice, N. Ahn for providing MEK plasmids and for sharing unpublished data, G. Gorbsky and M. Weber for sharing unpub-

Applicant's Copy 09/301,766

XP-002124106

Catalytic Plasticity of Fatty Acid Modification Enzymes Underlying Chemical Diversity of Plant Lipids

Pierre Broun,* John Shanklin,*† Ed Whittle, Chris Somerville†

Higher plants exhibit extensive diversity in the composition of seed storage fatty acids. This is largely due to the presence of various combinations of double or triple bonds and hydroxyl or epoxy groups, which are synthesized by a family of structurally similar enzymes. As few as four amino acid substitutions can convert an oleate 12-desaturase to a hydroxylase and as few as six result in conversion of a hydroxylase to a desaturase. These results illustrate how catalytic plasticity of these diiron enzymes has contributed to the evolution of the chemical diversity found in higher plants.

All higher plants contain one or more oleate desaturases that catalyze the O_2-dependent insertion of a double bond between carbons 12 and 13 of lipid-linked oleic acid ($18{:}1^{\Delta 9}$) to produce linoleic acid ($18{:}2^{\Delta 9,12}$) (*1*). In contrast, only 14 species in 10 plant families have been found to accumulate the structurally related hydroxy fatty acid, ricinoleic acid (D-12-hydroxyoctadec-*cis*-9-enoic acid) (*2*), which is synthesized by an oleate hydroxylase that exhibits a high degree of sequence similarity to oleate desaturases (*3*). The oleate desaturases and hydroxylases are integral membrane proteins, which are members of a large family of functionally diverse enzymes that includes alkane hydroxylase, xylene monooxygenase, carotene ketolase, and sterol methyloxidase (*1*). These nonheme iron–containing enzymes use a diiron cluster for catalysis (*4*) and contain three equivalent histidine clusters that have been implicated in iron binding and shown to be essential for catalysis (*1*). This class of proteins exhibits no significant sequence identity to the soluble diiron-containing enzymes which represent a similar diversity of enzymatic activities that include plant acyl-ACP desaturases, methane monooxygenase, propene monooxygenase, and the R2 component of ribonucleotide reductase (*1*, *5*). The catalytic activities of these enzymes has been mimicked by a synthetic diiron-containing complex with a coordination sphere composed entirely of nitrogen ligands (*6*).

The oleate hydroxylase from the crucifer *Lesquerella fendleri* has about 81% sequence identity to the oleate desaturase from the crucifer *Arabidopsis thaliana* and about 71% sequence identity to the oleate hydroxylase from *Ricinus communis* (*7*). The observation that these crucifer desaturase and hydroxylase enzymes are more similar than the two hydroxylases, and the presence of ricinoleic acid in a small number of distantly related plant species, suggests that the capacity to synthesize ricinoleate has arisen independently several times during the evolution of higher plants, by the genetic conversion of desaturases to hydroxylases.

Comparison of the amino acid sequences of the hydroxylases from *L. fendleri* and *R. communis* with the sequences for oleate desaturases from *Arabidopsis*, *Zea mays*, *Glycine max* (two sequences), *R. communis*, and *Brassica napus* revealed that only seven residues were strictly conserved in all of the six desaturases but divergent in both of the hydroxylases. The role of these seven residues was assessed by using site-directed mutagenesis to replace the residues found in the *Lesquerella* hydroxylase, LFAH12, with those from the equivalent positions in the desaturases (*8*, *9*). In a reciprocal experiment, we replaced the seven residues in the *Arabidopsis* FAD2 oleate desaturase with the corresponding *Lesquerella* hydroxylase residues (*10*). The activity of the modified and unmodified genes was then determined by expressing them in yeast and transgenic plants before analyzing the composition of the total fatty acids. Technical difficulties limited the utility of direct measurements of enzyme activity in cell extracts (*11*).

The mutant hydroxylase and desaturase genes containing all seven substitutions (designated m7LFAH12 and m7FAD2, respectively) were expressed in yeast cells under transcriptional control of the GAL1 promoter. Transgenic cells were harvested after induction and their total fatty acid composition determined by gas chromatography. Wild-type yeast cells do not accumulate detectable concentrations of diunsaturated or hydroxylated fatty acids (*12*). Expression of FAD2 caused the accumulation of about 4% diunsaturated fatty acids (16:2 and

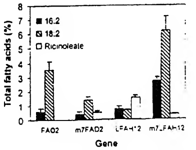


Fig. 1. Fatty acid composition of yeast cells expressing desaturase and hydroxylase genes. Cultures were induced in growth medium containing galactose, ~2 × 10^8 cells were harvested, and fatty acids were extracted and modified for analysis by gas chromatography, as described (*7*). Values are the averages (±SE) obtained from five cultures of independent transformants.

P. Broun and C. Somerville, Carnegie Institution of Washington, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA. J. Shanklin and E. Whittle, Biology Department, Brookhaven National Laboratory, Upton, NY 11973, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed.

18:2) but no detectable hydroxy fatty acids (Fig. 1). Expression of LFAH12 caused the accumulation of about 1.4% dunsaturated fatty acids and 1.5% ricinoleic acid, confirming the mixed function of this enzyme (7). Cells expressing m_1FAD2 accumulated ricinoleic acid to ~0.8% of total fatty acids and had ~50% reduction in the accumulation of dunsaturated fatty acids (Fig. 1). Thus, replacement of the seven residues (*10*) converted a strict desaturase to a bifunctional desaturase-hydroxylase comparable in activity to the unmodified *Lesquerella* hydroxylase.

The amount of desaturase activity of the LFAH12 enzyme is relatively low compared with its hydroxylase activity (7). However, yeast cells expressing LFAH12 accumulated linoleic and ricinoleic acids to similar concentrations, possibly because linoleic acid is more stable than ricinoleic acid in yeast cells. In cells expressing m_1FAH12, the ratio of 18:2 dunsaturated fatty acid to ricinoleic acid was, on average, 43 times that in cells expressing LFAH12. There was also a 16-fold increase in the ratio of 16:2 dunsaturated fatty acid to ricinoleic acid. Notwithstanding the quantitative limitations of the assay system, noted above, these results indicate a major increase in desaturase activity and a decrease in hydroxylase activity upon introduction of the seven desaturase-equivalent residues into LFAH12.

The activity of the mutant enzymes in planta was examined by using the corresponding genes to produce stable transgenic plants in an *Arabidopsis fad2* mutant, which is deficient in oleate desaturase activity (*13*). Expression of LFAH12 under transcriptional control of the constitutive cauliflower mosaic virus (CaMV) 35S promoter resulted in accumulation of high concentrations of hydroxy fatty acids in seeds (7), but no detectable suppression of the *fad2* mutant phenotype in leaves (Fig. 2). In contrast, expression of m_1LFAH12 under the same circumstances resulted in complete suppression of the *fad2* phenotype in 8 out of 10 transgenic plants analyzed (Fig. 2). There was an average 21-fold increase in the ratio of linoleate to oleate in leaf fatty acids and a small increase in the amount of linolenic acid. These results, which are consistent with the results of the yeast assays, confirm that expression of m_1LFAH12 in plants deficient in oleate desaturation has identical phenotypic consequences to expressing a wild-type desaturase such as FAD2 (*13*).

To evaluate the effect of the seven mutations on the activity of the gene encoding FAD2, we expressed FAD2 and m_1FAD2 in the *Arabidopsis fad2* mutant under the control of the strong seed-specific promoter from the *B. rapa* napin gene. As expected from previous studies (7), none of the 15 transgenic lines expressing the FAD2 gene accumulated detectable hydroxy fatty acids, although the ratio of linoleate to oleate accumulation was increased an average of 10-fold as compared with untransformed controls. In the transgenic lines expressing m_1FAD2, the amount of hydroxylated fatty acids, which included ricinoleic, densipolic, and lesquerolic acids, composed up to 9.4% of total seed fatty acids (Fig. 3). The ratio of seed linoleate to oleate contents was increased an average of 6.4-fold (*14*), which indicated that m_1FAD2 exhibited significant desaturase activity, albeit less than the wild-type FAD2 gene. The high concentrations of hydroxy fatty acid accumulation observed in transgenic plants expressing m_1FAD2 indicated that the modified desaturase had comparable levels of hydroxylase activity, in the in planta assay, to the native *Lesquerella* hydroxylase enzyme.

To determine whether any single amino acid residue of the seven had a major effect on the ratio of hydroxylase to desaturase activities, we introduced each of the seven FAD2-equivalent residues (*8*) individually into the LFAH12 enzyme. None of the enzymes containing single amino acid substitutions had activities that differed significantly from the wild-type hydroxylase enzyme when expressed in yeast (*14*). We also tested seven modified LFAH12 genes containing all combinations of six desaturase-equivalent residues (Fig. 4). Each of the seven constructs produced a ratio of dunsaturated to hydroxylated fatty acids that was similar to the ratio produced by the m_1FAH12 enzyme. Thus, as few as six residues principally determine the ratio of desaturation or hydroxylanon activity. All lines showed somewhat reduced levels of desaturase activity, with the largest reductions of ~40% seen in F218Y and G105A. Therefore, we made a construct in which both these changes were combined (xF218Y G105A). This construct exhibited similar activity to the individual F218Y and G105A mutants (*14*), suggesting that their effects are redundant and that the observed changes in activity result from interactions of more than two of the seven residues. Considered together, these results indicate that no single amino acid position plays an essential role in catalytic outcome. Rather, changes in activity result from a combined effect of several amino acid positions that have partially overlappng effects.

Because four of the seven amino acids are adjacent to histidine residues that have been identified as essential to catalysis (*1*), we hypothesized that these four residues may be of greatest importance to the outcome of the reaction. A modified FAD2 enzyme, designated m_2FAD2, was constructed in which these four amino acids

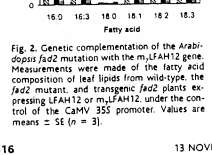


Fig. 2. Genetic complementation of the *Arabidopsis fad2* mutation with the m_1LFAH12 gene. Measurements were made of the fatty acid composition of leaf lipids from wild-type, the *fad2* mutant, and transgenic *fad2* plants expressing LFAH12 or m_1LFAH12 under the control of the CaMV 35S promoter. Values are means ± SE (n = 3).

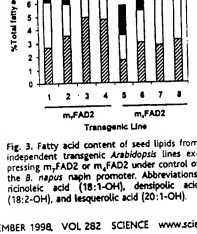


Fig. 3. Fatty acid content of seed lipids from independent transgenic *Arabidopsis* lines expressing m_1FAD2 or m_2FAD2 under control of the *B. napus* napin promoter. Abbreviations: ricinoleic acid (18:1-OH), densipolic acid (18:2-OH), and lesquerolic acid (20:1-OH).

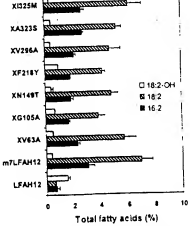


Fig. 4. Contribution of individual amino acid substitutions to the activity of the modified *Lesquerella* hydroxylase. Seven derivatives of the m_1LFAH12 gene containing all combinations of six out of seven substitutions were introduced into yeast cells, and the fatty acid composition of five independent cultures was measured. The "X" designation refers to the unmodified amino acid (that is, enzyme X1325M contains all of the seven substitutions except 1325M).

were replaced by their equivalents from the *Lesquerella* hydroxylase (T148N, A296V, S322A, M324I). Expression of m_4FAD2 in seeds of wild-type *Arabidopsis* resulted in the accumulation of average concentrations of hydroxy fatty acids that were similar to those obtained with m-FAD2 (Fig. 3). Thus, only four changes are required to convert a strict desaturase to an enzyme that retains some desaturase activity but is also an efficient hydroxylase.

Biochemical and structural similarities between the desaturase and hydroxylase, in addition to recent kinetic isotope experiments, suggest that there is a common initial oxidation event at C-12 for both enzymes (*15*). Thus, it seems likely that the different functional outcomes represent a partitioning between two reaction pathways that diverge after initial C-12 hydrogen abstraction such that one pathway favors a second hydrogen abstraction whereas the other favors oxygen transfer. We envision that because no specific single amino acid change is required, and in view of the substantial effect of the four residues that abut the active site histidines, the differences between desaturase and hydroxylase outcome is influenced by changes in active site geometry. Examples of such changes might include the relative positioning of the substrate with respect to the iron center, the coordination geometry of the iron ions, or the active site hydrogen bonding network. Whatever the case, this mode of evolving new catalytic activity differs from the more general case in which the evolution of new activities involves the incorporation of new catalytic groups into the active site (*16*).

Acetylenic and epoxy fatty acids are produced by desaturation and epoxidation of double bonds by enzymes that are structurally similar to the enzymes described here (*17*). Thus, variations of the same catalytic center can catalyze the formation of at least four different functional groups in fatty acids. Because various combinations of these four functional groups define most of the chemical complexity found among the hundreds of different fatty acids that occur in higher plants (*2*), it is now apparent that most of the chemical complexity of plant fatty acids can be accounted for by divergence of a small number of desaturases. Extrapolating from the results described here, it also seems very likely that a small number of amino acid substitutions will account for the functional divergence of desaturases, hydroxylases, expoxgenases, and acetylenic bond-forming enzymes.

References and Notes

1. J. Shanklin and E. B. Cahoon, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 49, 611 (1998).
2. F. N. van de Loo, B. Fox, C. R. Somerville, in *Plant Lipids*, T. Moore, Ed. (CRC Press, Boca Raton, FL, 1993), pp. 91–126.
3. R. A. Moreau and P. K. Stumpf, *Plant Physiol.* 67, 672 (1981); F. N. van de Loo, P. Broun, S. Turner, C. R. Somerville, *Proc. Natl. Acad. Sci. U.S.A.* 92, 6743 (1995).
4. J. Shanklin, C. Achim, H. Schmidt, B. G. Fox, E. Munck, *Proc. Natl. Acad. Sci. U.S.A.* 94, 2981 (1997).
5. L. Que Jr., *Science* 253, 273 (1991); J. Shanklin and C. R. Somerville, *Proc. Natl. Acad. Sci. U.S.A.* 88, 2510 (1991); M. Suzuki, T. Hayakawa, J. P. Shaw, M. Rekik, S. Harayama, *J. Bacteriol.* 173, 1690 (1991); Y. Lindqvist, W. Huang, G. Schneider, J. Shanklin, *EMBO J.* 15, 4081 (1996); J. D. Pikus *et al.*, *Biochemistry* 35, 9106 (1996).
6. C. Kim, Y. Dong, L. Que, *J. Am. Chem. Soc.* 119, 3635 (1997).
7. P. Broun, S. Boddupalli, C. R. Somerville, *Plant J.* 13, 201 (1998).
8. The seven mutations were V63A, G105A, N149T, F218Y, V296A, A323S, and I325M numbered relative to the LFAH12 sequence, where the first residue at the indicated position was mutated to the second residue. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
9. Mutagenic oligonucleotides were used to introduce nucleotide substitutions into cloned genes by overlap-extension polymerase chain reaction (PCR) [W. Ito, H. Ishiguro, Y. Kurosawa, *Gene* 102, 67 (1991)]. In a first step, we amplified overlapping fragments in separate PCR reactions using primer pairs designed to introduce mutations. The products were gel purified, then assembled in a PCR reaction primed with terminal primers only. Modified LFAH12 genes containing one or seven substitutions were constructed with pLFAH12-1 as template (*7*) and primers mH1 through mH5 or mH67. Modified LFAH12 genes containing only six mutations were constructed with m_7LFAH12 as a template and one of the primers H1 to H7 to revert one of the mutations. The 5' end of terminal primers was modified to allow the introduction of convenient restriction sites for the cloning of PCR products. The m_7FAD2 was constructed with oligonucleotides D1 to D5 and D67. The PCR conditions were as follows: 10 ng of plasmid DNA, 200 μM deoxynucleotide triphosphates, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 200 mM tris-HCl (pH 8.8), 20 mM $MgSO_4$, 1% (v/v) Triton X-100, bovine serum albumin (1000 μg/ml), 3 mM $MgCl_2$, 5% (v/v) dimethyl sulfoxide, 125 pmol of each primer, 1.25 U of Pfu polymerase (Stratagene), to a final volume of 50 μl. Amplifications conditions were as follows: 4 min denaturation step at 94°C, followed by 30 cycles of 92°C for 1 min, 50°C for 1 min, and 72°C for 2 min concluded with a final extension step at 72°C for 5 min. PCR products were purified from agarose or polyacrylamide gels. For the second PCR step, 10 ng of purified overlapping fragments were used as templates in PCR reactions as above except that only 15 cycles were used. PCR fragments encoding modified LFAH12 enzymes were cloned into pLFAH12-1 cut with Pst I and one of Sma I or Eco RV or Sac I. All inserts were sequenced. Yeast expression vectors containing wild-type or modified LFAH12 genes were constructed by excising inserts from the above constructs with the enzymes Hind III and Sac I and cloning them into the Hind III–Sac I sites of pYESII (Invitrogen). Constructs for plant transformation were made by cloning the Stu I–Sac I fragment from modified LFAH12 genes into the Sma I–Sac I sites of pBI121. The FAD2 cDNA clone 146M12T7 was amplified with Pfu DNA polymerase by using primers D5' and D3' to introduce restriction sites for Kpn I and Sma I immediately upstream of the initiation codon, and Sac I and Eco RI restriction sites following the terminator codon. The fragment was cloned into the Eco RV site in the vector pZErO (Invitrogen). For expression of FAD2 in yeast, the insert was excised by restriction with Kpn I and Eco RI and cloned into the corresponding sites in pYESII, resulting in plasmid pYESII-F2. Binary Ti-vector pDN was constructed for seed-specific expression of FAD2 genes. In a first step, the napin promoter was amplified from rapeseed DNA with primers ggcgtcgacaagcttctgcggatcaagcagctttca and ggttttgagtagtgatgtcttgtatgttctagatggtaccgtac. A Hind III–Bgl II fragment was cloned into the Hind III–Bgl II sites of pBI121 (Clontech), replacing the 35S promoter. FAD2 coding sequences were excised from pYESII-F2 with Sma I and Sac I and cloned into pDN with corresponding restriction sites. The construction of the m_7FAD2 cDNA encoding a modified FAD2 enzyme containing seven amino acid substitutions was achieved with overlap extension PCR. After the second round of assembly-amplification with the primers D5' and D3' the PCR products were treated exactly as the amplified wild-type FAD2 sequence described above.

   Plant expression constructs were introduced into *Agrobacterium tumefaciens* strain GV3101 pMP90 by electroporation and used to transform *Arabidopsis fad2* mutant plants by vacuum infiltration [D. Bouchez, C. Camilleri, M. Caboche, *C. R. Acad. Sci. Ser. III* 316, 1188 (1993)]. The oligonucleotides used were mH1f atcactttagcttcttgcttct, mH1r agaagcaagaagctaaagtgat, mH2f ctgggcattgcccatgaatgtggtcacc, mH2r ggtgaccacattcatgggcaatgacccag, mH3f caccattccaacactggatcctagaa, mH3r ttctaggatcccagtgttggaatggtg, mH4f catgcacctatctataaggaccgtg, mH4r cacggtccttatagataggtgcatg, mH5f agaggagctttggctacggtagac, mH5r gtctaccgtagccaaagctcctct, mH67f catctctttcaactatgccgcatt, mH67r aatgcggcatagttgaeaaagagatg, mH6f catctctttcaactataccgcatt, mH6r aatgcggtatagttgaaaagagatg, mH7f catctctttgcaactatgccgcatt, mH7r aatgcggcatagttgcaaagagatg, H1f atcactttagtttcttgcttct, H1r agaagcaagaaactaaagtgat, H2f ctgggcattggccatgaatgtggtcacc, H2r ggtgaccacattcatggccaatgacccag, H3f caccattccaacaatggatccctagaa, H3r ttctagggatccattgttggaatggtg, H4f catgcacctatctttaaggaccgtg, H4r cacggtccttaaagataggtgcatg, H5f agaggagctttggttacggtagac, H5r gtctaccgtaaccaaagctcctct, H6f catctctttgcaactataccgcatt, H6r aatgcggtatagttgcaaagagatg, H7f catctctttgcaactatac cgcatt, H7r aatgcggtatagttgcaaagagatg, mD1f gacatcattatagtctcatgcttctact, mD1r agtagaagcatgagactataatgatgtc, mD2f ctgggtcataggccacgaatgcggtc, mD2r gaccgcattcgtggcctatgacccag, mD3f caccattccaacaatggatccctcgaa, mD3r ttcgagggatccattgttggaatggtg, mD4f ccccaacgctcccatcttcaatgaccgaga, mD4r tctcggtcattgaagatgggagcgttgggg, mD5f caggggagctttggttaccgtagacagag, mD5r ctctgtctacggtaaccaaagctcccctg, mD67f cacctgttcgcgacaataccgcattataacgc, mD67r gcgttataatgcggtattgtcgcgaacaggtg, H5' tatcgaaggcctgatgggtgct, H3' ctcgcagtatcgagctcataacttattgtt, D5 gatcggtacccgggatgggtgcaggtggaagaatgccgg, and D3 gatcgaattcgagctctcataacttattgttgtaccagtacacacc. The bold and underline letters indicate altered nucleotides and codons, respectively.
10. The seven mutations were A63V, A104G, T148N, Y217F, A295V, S322A, and M324I, designated on the basis of the numbering of the *Arabidopsis* FAD2 sequence.
11. The enzymes are integral membrane proteins that act on fatty acids esterified to lipids and require cytochrome b_5 reductase and cytochrome b_5 for activity. The difficulty of quantitatively incorporating labeled lipids into isolated membranes and ensuring that cytochrome b_5 and b_5 reductase are not limiting restricts the utility of direct measurements of enzyme activity. Our best estimates of oleate desaturase or oleate hydroxylase activities in crude microsomal preparations from *Arabidopsis* roots indicated specific activities of 1.2 and 0.3 pmol/mg of protein per minute, respectively.
12. P. S. Covello and D. W. Reed, *Plant Physiol.* 111, 223 (1996); S. Kajiwara *et al.*, *Appl. Environ. Microbiol.* 62, 4309 (1996).
13. M. Miquel and J. Browse, *J. Biol. Chem.* 267, 1502 (1992).
14. P. Broun *et al.*, data not shown.
15. P. H. Buist and B. Behrouzian, *J. Am. Chem. Soc.* 120, 871 (1998).
16. P. C. Babbitt and J. A. Gerlt, *J. Biol. Chem.* 272, 30591 (1997).
17. M. Lee *et al.*, *Science* 280, 915 (1998).

18. We thank S. Stymne for communicating results before publication, and B. Behrouzian for helpful discussion. This work was supported in part by grants from the U.S. Department of Energy Office of Basic Energy Sciences to C.R.S. (DE-FG02-97ER20133), and to J.S. and E.J.W. The Monsanto Company awarded P.B. a fellowship.

31 July 1998; accepted 9 October 1998

Eur. J. Biochem. 38, 103–110 (1973)

The Function of *myo*-Inositol in the Biosynthesis of Raffinose

Purification and Characterization of Galactinol:Sucrose 6-Galactosyltransferase from *Vicia faba* Seeds

Ludwig Lehle and Widmar Tanner

Fachbereich Biologie der Universität Regensburg

(Received April 18/June 28, 1973)

1. An enzyme from *Vicia faba* seeds is described which transfers the galactosyl moiety of galactinol to sucrose giving rise to raffinose and *myo*-inositol.

2. The enzyme was purified about 400-fold through 6 steps. A molecular weight of 80000 has been determined by gel-filtration and of 100000 by glycerol density gradient centrifugation.

3. The enzyme galactinol:sucrose 6-galactosyl transferase is different from α-galactosidase; these two activities as well as the stachyose-synthesizing enzyme separate during purification.

4. The transferase showed a high acceptor specificity. Out of 10 acceptors tested a transfer only to sucrose took place. This transfer was 5 times faster than the hydrolysis of galactinol. Galactinol, *p*-nitrophenyl-α-D-galactopyranoside and raffinose, but not UDP-galactose, could act as donors.

5. The enzyme catalyzes an exchange reaction between raffinose and [^{14}C]sucrose. This partial reaction is less sensitive towards heat inactivation and SH-poisons than the total reaction.

6. The pH-optimum of the reaction was found to be pH 7.0, the temperature optimum 42 °C. Heat inactivation could be prevented to some extent by galactinol and raffinose. In the presence of 0.4 mM sucrose the K_m-value for galactinol was 7 mM and for raffinose 10 mM. For sucrose a K_m-value of 1 mM in the synthesis reaction has been determined.

7. The transferase activity is high enough to explain the synthesis rate *in vivo* of all the raffinose-type sugars present in the seeds.

8. The physiological meaning of the results as well as the metabolic function of *myo*-inositol is discussed.

One of the major exceptions to Leloir's mechanism [1] of glycosidic linkage formation in nature has been discovered in the biosynthesis of a group of plant oligosaccharides, the sugars of the raffinose family [2, 3]. Besides sucrose these sugars are the most common and widespread ones in higher plants and have a function as storage and transport material [4–6]. Whereas evidence *in vivo* and *in vitro* [2, 7–9] has firmly established that the biosynthesis of stachyose and verbascose proceeds *via* a transglycosylation of the galactosyl-moiety from galactinol [L-1-(*O*-α-D-galactopyranosyl)-*myo*-inositol] to raffinose and stachyose, respectively [Eqns (3) and (4) below], conflicting evidence has been published concerning the biosynthesis of raffinose, the smallest member of the homologous series of these oligosaccharides.

On the one hand evidence for the reaction sequence (1) and (2), analogous to stachyose and verbascose synthesis, has been presented [10]. On the other hand a transfer of the galactosyl moiety from UDP-galactose to sucrose has

UDP-galactose + *myo*-inositol → galactinol + UDP (1)

Galactinol + sucrose ⇄ raffinose + *myo*-inositol (2)

Galactinol + raffinose ⇄ stachyose + *myo*-inositol (3)

Galactinol + stachyose ⇄ verbascose + *myo*-inositol (4)

Abbreviation. Gal-αONp, *p*-nitrophenyl-α-D-galactopyranoside.

Trivial Name. Galactinol, L-1-(*O*-α-D-galactopyranosyl)-*myo*-inositol.

Enzymes. α-Galactosidase or α-D-galactoside galactohydrolase (EC 3.2.1.22); galactinol:raffinose 6-galactosyltransferase (EC 2.4.1.–); aldolase or fructose-1,6-bisphosphate D-glyceraldehyde-3-phosphate lyase (EC 4.1.2.13).

been reported [11—13]. However, in this case the enzyme preparations were fairly crude and the possibility cannot be excluded that the sum of reaction (1) and (2) has been measured. Reaction (1) has been originally described by Frydman and Neufeld [14].

In the report to follow a 400-fold purification of the galactinol: sucrose 6-galactosyl transferase, the enzyme catalyzing reaction (2), from *Vicia faba* seeds will be described. The enzyme also catalyzes an exchange reaction between raffinose and sucrose, which is considerably more stable than the reaction responsible for net synthesis of raffinose. This latter observation explains the fact that Moreno an Cardini [15] have been able to observe only the exchange reaction in wheat germ extracts.

MATERIALS AND METHODS

Purification Procedure

All procedures were carried out at about 4 °C.

Step 1. Preparation of Crude Extract. 200 g ripe seeds from *Vicia faba* were powdered in a Waring Blendor and then extracted in a chilled mortar in two portions each with 200 ml of 0.1 M Tris-HCl buffer pH 7.3 containing dithioerythritol 5 mM. The homogenate was centrifuged for 30 min at $27000 \times g$ giving a clear supernatant of about 250 ml.

Step 2. Treatment with Protamine Sulfate. The supernatant was brought to a protein concentration of 50 mg/ml with the same buffer as used for step 1. A 2% protamine sulfate solution was added to a final ratio of 9 mg protamine sulfate per 100 mg protein. After 30 min of stirring, the resulting precipitate was centrifuged off and discarded.

Step 3. Ammonium Sulfate Fractionation. To the protamine-treated supernatant saturated, cold ammonium sulfate solution, pH 7.3, was slowly added with continuous stirring to give 33% saturation. After 30 min, the precipitate was separated by centrifugation and the supernatant was brought to 55% saturation. The pellet obtained after centrifugation was dissolved in 70 ml 0.1 M Tris-HCl pH 7.3 containing 5 mM dithioerythritol and dialyzed overnight against 3 l of 0.05 M Tris-HCl pH 7.5, containing 1 mM dithioerythritol.

Step 4. Column Chromatography on DEAE-Cellulose. The dialyzed enzyme solution was adsorbed on a DEAE-celluloso column (2.5×30 cm) which had been equilibrated with 0.01 M Tris-HCl pH 7.5 containing 0.05 M KCl and 1 mM dithioerythritol. After the column was washed with equilibration buffer until all protein not bound was removed, 1 l linear gradient of 0.05 M KCl to 0.2 M KCl in 0.01 M Tris-HCl with 1 mM dithioerythritol was used for elution. Fractions of 6 ml were collected and those with the highest specific activity were pooled and concentrated to a small volume in an Amince ultrafiltration cell with filter No XM-50.

Step 5. Sephadex G-200 Gel Chromatography. The pooled and concentrated fractions were loaded onto a column (2.5×80 cm) of Sephadex G-200, equilibrated with 0.01 M Tris-HCl buffer pH 7.5 containing 0.1 M KCl and 2 mM dithioerythritol. The column was eluted at a flow rate of 4 ml/h; 2-ml fractions were collected and the active fractions (100—120) were pooled and concentrated as described before.

Step 6. Hydroxyapatite Chromatography. After dialysis against 0.01 M Tris-HCl with 2 mM dithioerythritol pH 7.5, the enzyme solution was applied to a column (2.5×13 cm) of hydroxyapatite, which had been equilibrated with 0.01 M potassium phosphate buffer pH 7.5 containing 2 mM dithioerythritol. Elution was carried out stepwise with 100 ml potassium phosphate buffer of the following concentrations: (a) 0.01 M; (b) 0.05 M; (c) 0.1 M; (d) 0.2 M. The enzyme was eluted with 0.2 M buffer. The active fractions were again concentrated as described above.

Tests for Enzymic Activities

Galactosyltransferase: Synthesis and Exchange Reaction. Two tests have been used, to measure the transfer of the galactosyl moiety from galactinol to sucrose. In test I the amount of [^{14}C]raffinose formed from [^{14}C]sucrose has been determined. The incubation mixture contained in a total volume of 50 µl: 5 µmol Tris-HCl pH 7.2, 1 µmol galactinol, 0.02 µmol [^{14}C]sucrose (35 µCi/µmol) and enzyme. After incubation of 1—4 h at 32 °C the reaction was stopped with 0.2 ml ethanol and the preparation was centrifuged; the supernatant fluid was separated on Whatman No 1 in the solvent system n-butanol—pyridine—water—acetic acid (60:40:30:3, v/v/v/v). Radioactive spots were located with a strip scanner, cut out, and measured directly on paper in a scintillation counter in toluene—2,5-diphenyloxazole (efficiency 70%). This test was also applied for the exchange reaction with the only exception that 0.5 µmol raffinose was used instead of galactinol. The linear relationship between product formation, protein concentration up to 4 mg and incubation time up to 6 h has already been demonstrated for both reaction [10] and has since also been shown to be valid for the more purified enzyme preparations used in the kinetic experiments.

Test II is based on the galactosyl transfer from ^{14}C-labelled galactinol to sucrose. With this test one can study in addition the amount of galactose set free by the hydrolyzing activity of the transferase. The incubation mixture contained in a total volume of 50 µl: 5 µmol Tris-HCl pH 7.2, 0.013 µmol

[^{14}C]galactinol (7 μCi/μmol), 0.5 μmol sucrose and enzyme. The chromatographic separation was carried out in a solvent system of α-picoline—ammonia—water (70:28:2, v/v/v) until the front had reached half way down the paper. Then a second run in the solvent system *n*-butanol—pyridine—acetic acid—water (60:40:3:30, v/v/v/v) followed. Other conditions were the same as in test 1.

α-Galactosidase. The enzyme was assayed by following the initial rate of *p*-nitrophenyl-α-D-galactopyranoside (Gal-αONp) hydrolysis. Enzyme solution was incubated at 32 °C with 25 μmol potassium phosphate buffer pH 5.5 and 0.0 μmol Gal-αONp for 15 min. The reaction was stopped by adding 5.0 ml of cold 0.1 M Na_2CO_3 and the yellow colour of *p*-nitrophenol was measured at 405 nm. Controls with Gal-αONp as well as with protein alone were run concurrently and all values appropriately corrected.

Determinations of Molecular Weight

The molecular weight was determined on a Sephadex G-200 column (2.5 × 80 cm) according to Andrews [16]. The column was eluted with 0.01 M Tris-HCl pH 7.5 containing 0.1 M KCl and 2 mM dithioerythritol. The calibration was obtained by determination of the elution volumes of a number of reference proteins of known molecular weight. The sedimentation constant of the enzyme was determined by centrifugation through a linear 5-ml gradient ranging from 5—20% glycerol in 0.05 M Tris-HCl pH 7.5 containing 5 mM dithioerythritol. The samples were centrifuged in the SWL 50 rotor of a Spinco L 2-65 B for 14 h at 0 °C. Then the tubes were punctured and fractions of 3 drops collected with the aid of a fraction collector. As reference protein aldolase was used.

Polyacrylamide-Gel Electrophoresis

The purity of the various purification steps was routinely checked by polyacrylamide gel electrophoresis in a 7.5% acrylamide gel according to Maurer [17]. Electrophoresis was performed at 2.0 mA/tube until the bromphenolblue band had reached the bottom of the tube. Fixation and staining were carried out according to Chrambach *et al.* [18].

Other Procedures

Protein determinations were carried out according to Lowry *et al.* [19] with bovine serum albumin as a standard. Labelled galactinol was isolated by paper chromatography from the water-soluble extract of lamium leaves after photosynthesis in $^{14}CO_2$ according to Kandler [20]. A sample of unlabelled galactinol was generously supplied by Dr R. M. McCready (USDA, Agricultural Research Service, Albany).

RESULTS

Purification of Galactinol: Sucrose 6-Galactosyltransferase

Table 1 summarizes the results of the overall purification. Starting from a crude extract which a specific activity of 0.071 nmol × mg^{-1} × h^{-1} a preparation was obtained with a specific activity of 29.8 nmol × mg^{-1} × h^{-1} (peak II of hydroxyapatite chromatography). The results show that the enzyme catalyzing the synthesis of stachyose [8] separates

Table 1. *Purification procedure for galactinol : sucrose 6-galactosyltransferase*
Figures in brackets represent percentage of original activity. Stachyose was measured as described previously [7]

Fraction	Total protein	Raffinose synthesis		Exchange reaction		Stachyose synthesis		α-Galactosidase activity	
		Total activity	Specific activity	Total activity	Specific activity	Total activity	Specific activity	Total activity	Specific activity
	mg	nmol/h	nmol × h^{-1} × mg^{-1}	nmol/h	nmol × h^{-1} × mg^{-1}	nmol/h	nmol × h^{-1} × mg^{-1}	μmol/h	μmol × h^{-1} × mg^{-1}
1. Crude extract	22000	1576 (100)	0.071	1980 (100)	0.091	2148 (100)	0.098	1540 (100)	70
2. Protamine sulfate	9350	1750 (110)	0.187	2072 (105)	0.222	2306 (107)	0.246	272 (18)	29
3. Ammonium sulfate	2688	1398 (88)	0.520	1730 (88)	0.647	1488 (69)	0.553	161 (10)	58
4. DEAE-cellulose	182	557 (35)	3.067	614 (31)	3.376	0 (0)	0	5 (0.3)	29
5. Sephadex G-200	50	262 (17)	5.273	359 (18)	7.108	0 (0)	0	0 (0)	0
6. Hydroxyapatite									
Peak I	4.5	3.4 (0.2)	0.703	32.9 (1.6)	7.326	0 (0)	0	0 (0)	0
Peak II	2.3	68.5 (4.3)	29.800	58.5 (2.9)	25.345	0 (0)	0	0 (0)	0

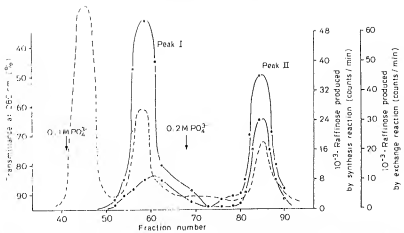


Fig. 1. *Hydroxyapatite chromatography of galactinol : sucrose 6-galactosyltransferase.* Suitable aliquots were tested by test I for synthesis activity (● ● ●) and exchange activity (○ ○ ○); absorbance at 280 nm (– – –)

from the corresponding raffinose-synthesizing enzyme In addition it has to be pointed out that the purified enzyme is different from an α-galactosidase, since the hydrolyzing activity towards *p*-nitrophenyl-α-D-galactopyranoside (Gal-αONp), known to be a good substrate for α-galactosidases, separates likewise from the raffinose-synthesizing activity. Since the galactinol : sucrose 6-galactosyl transferase is the most labile of the plant galactosyl transferases known (*e.g.* [10]), it seems unlikely that an inactivation instead of a separation of the other two enzymes had occurred during the purification. The considerable decrease of the α-galactosidase activity in step 2 may on the other hand be the reason for the observed increase of the total raffinose-synthesizing activity in this fraction, since less of the newly synthesized raffinose will be lost by hydrolysis.

The preparation from *Vicia faba* also catalyzes an exchange reaction between raffinose and sucrose according to the following equation:

Raffinose + [^{14}C]sucrose
$\rightleftarrows$ [^{14}C]raffinose + sucrose.

This reaction has originally been described by Moreno and Cardini [15]; their enzyme preparation from wheat germ, however, did not catalyze the synthesis of raffinose. Through all the steps given in Table 1 (except for step 6; see Discussion below) the exchange reaction parallels the synthesis activity. Thus both reactions most likely are catalyzed by one and the same enzme.

In the last purification step two active transferase peaks (I and II in Fig. 1) were obtained. The main fraction, peak II, was eluted with a buffer concentration of 0.2 M. Peak I, which had much lower specific activity, appeared at 0.1 M. Both fractions were able to catalyze the synthesis as well as the exchange reaction, although at different relative rates. Whereas peak I catalyzes the exchange reaction about 10 times faster than the synthesis of raffinose, peak II catalyzes the exchange reaction only at 85% the rate of synthesis reaction. Further experiments indicated that peak I is a modified form of the enzyme, which has lost most of its raffinose-synthesizing activity and shows a different elution behaviour as compared to the native enzyme. Thus, when peak II was chromatographed a second time on hydroxyapatite, again an active peak I and II was obtained. The observation made previously, that the activity for raffinose synthesis is lost more readily than the activity of the exchange reaction [10], is in agreement with the above finding.

When checked for purity by polyacrylamide gel electrophoresis the 400-fold purified fraction was not yet homogeneous; one major and three minor bands have been observed (Fig. 2). Although a strong attempt has been made to correlate the enzyme activity with one of the bands, this has failed; the enzymic activity always got lost during gel-electrophoreses, even in the presence of a variety of protecting agents.

The enzyme remained in the supernatant when the enzyme solution was centrifuged at 100000×*g* for 1 h.

Determination of Molecular Weight

The molecular weight of the enzyme was determined by two different methods. From the sedimentation profile in a glycerol density gradient a molecular weight of 100000 was obtained when compared to the sedimentation of aldolase (Fig. 3). With Sephadex G-200 gel chromatography on a standardized column (Fig. 4) a value of 80000 was

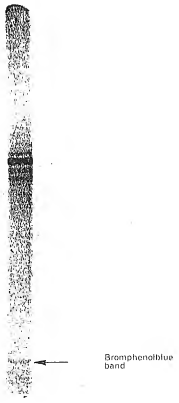


Fig. 2. *Disc-gel electrophoresis of peak II of the hydroxyapatite column*

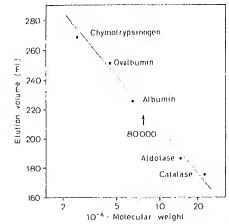


Fig. 4. *Determination of the molecular weight by Sephadex G-200 gel filtration.* The procedure is described in the text. The arrow indicates the position of the enzyme. Molecular weight is plotted on a log scale

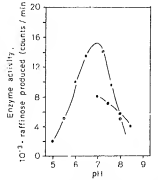


Fig. 5. *pH dependence of galactinol : sucrose 6-galactosyltransferase.* Assays were performed with potassium phosphate buffer (○ ○) and Tris-HCl buffer (● ●)

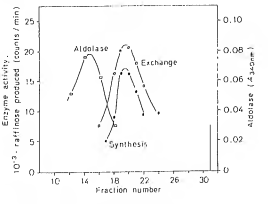


Fig. 3. *Sedimentation profile of galactinol : sucrose 6-galactosyltransferase in a 5—20% glycerol density gradient.* 100 µg of purified enzyme (Sephadex fraction) and 500 µg of aldolase were centrifuged for 16 h at 40000 rev./min. Enzyme activity has been tested for synthesis reaction (●——●) and exchange reaction (○ - - - ○). Aldolase served as a marker

determined. In each case, however, the same values were observed, whether synthesis or exchange activity had been tested.

Stability

When stored at 4 °C the crude extract lost 50% of its original activity in the synthesis reaction and 30% in exchange reaction within 3 days. The activity of the purified enzyme when frozen was unchanged for at least a month.

pH Optimum

The enzyme showed an optimum around pH 7.0. In the presence of potassium phosphate buffer the activity was higher than in the presence of Tris-HCl buffer (Fig. 5).

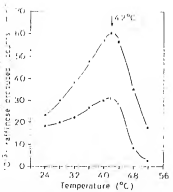


Fig. 6. *Effect of temperature on the activity of galactinol : sucrose 6-galactosyltransferase (Sephadex fraction).* (● ●) Synthesis reaction; (○ ○) exchange reaction

Table 2. *Effect of substrates on heat inactivation of galactinol : sucrose 6-galactosyltransferase*

140 µg enzyme (Sephadex fraction) was preincubated with or without substrate at 50 °C for 10 min. Then the rest of the incubation mixture was added and the test was carried out under standard conditions for 150 min

Substrate	Enzyme activity after preincubation			
	Synthesis reaction		Exchange reaction	
	counts/min	%	counts/min	%
Control (without heat)	22755	100	46258	100
– substrates	4350	19	16610	36
+ donor[a]	7507	33	29605	64
+ acceptor[b]	4515	20	17994	38

[a] Galactinol and raffinose.
[b] Sucrose.

Effect of Temperature on the Enzyme Activity

Fig. 6 shows the temperature profile of the enzyme activities. Maximum rate for both reactions occurs at 42 °C with a sharp drop beyond 44 °C, the synthesis reactions being somewhat more sensitive than the exchange reaction. In this connection it has been observed that galactional and raffinose prevent to some extent inactivation by heat (Table 2). Sucrose, however, at the concentrations used had no effect.

Inhibition with Sulfhydryl-Specific Reagents

One of the main reasons that the first step in the biosynthesis of raffinose sugars escaped detection for a rather long time has certainly been the requirement of the enzyme for strong SH-protecting agents [10]. This is especially true for the synthesis reaction. The different susceptibility of synthesis and exchange reaction is also reflected by the inhibition of the enzyme with iodoacetamide and *N*-ethylmaleimide (Table 3). The heavy metal ions Ag^{+}, Hg^{2+}, Zn^{2+} and Al^{3+} at a concentration of 1 mM inhibited the synthesis reaction of the enzyme to 100%; Mn^{2+} inhibited to 60%.

Table 3. *Inhibition of synthesis and exchange reaction by thiol-group specific reagents*

150 µg enzyme (Sephadex fraction) was incubated under standard conditions for 2 h. The inhibitor concentration was 1 mM

Inhibitor	Synthesis reaction		Exchange reaction	
	counts/min	(%)	counts/min	(%)
Control	21510	(0)	35690	(0)
Iodoacetamide	10995	(49)	31510	(11)
N-Ethyl-maleimide	2153	(90)	25490	(16)

Enzyme Kinetics

K_m values for galactinol, sucrose and raffinose have been determined (Fig. 7, Fig. 8 and Table 4). The Michaelis constant for sucrose was found to be 1 mM in the synthesis reaction and 2.9 mM in the exchange reaction in the presence of 0.02 M galactinol and raffinose, respectively. When the galactinol and raffinose concentrations were decreased 100-fold, the K_m for sucrose in the synthesis reaction stayed the same (1.4 mM). It was, however, considerably lower (0.47 mM) in the exchange experiment. This is consistent with the assumption that the binding site for raffinose and sucrose might be identical; a high raffinose concentration would then act as competitive inhibitor. On the other hand the sites for galactinol and sucrose seem to be different; a change in the concentration of galactinol has no influence on the K_m of sucrose. It has to be pointed out that the K_m-values for galactinol and raffinose given in Table 4 are only valid for a sucrose concentration of 0.4 mM.

Acceptor and Donor Specificity

The acceptor specificity has been tested by measuring the transfer of the ^{14}C-labelled galactosyl moiety from [^{14}C]galactinol to various acceptors. Out of 10 acceptors tested only a transfer to sucrose could be observed (Table 5). The purified enzyme cannot catalyze the biosynthesis of stachyose and verbascose. Both these enzymic activities have already been found in seeds from *Vicia faba* [8]. It should be noted that during the incubation of [^{14}C]galactinol some free [^{14}C]galactose was obtained due to the hydrolysis of galactinol. However, in the presence of sucrose the amount of galactose transferred was nearly 5 times greater than the amount of galactinol hydrolyzed (Table 5). In the absence of any acceptor considerable more galactose was set free. This can be interpreted as a competitition of sucrose with water. As donors only galactinol,

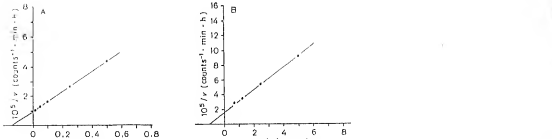


Fig. 7. *Lineweaver-Burk plots: raffinose-synthesis reaction.* (A) Galactinol in the presence of 0.4 mM sucrose; (B) sucrose in the presence of 0.01 M galactinol

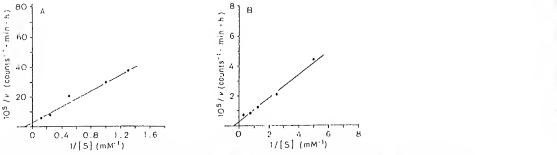


Fig. 8. *Lineweaver-Burk plots: exchange reaction.* (A) Raffinose in the presence of 0.4 mM sucrose; (B) sucrose in the presence of 0.01 M raffinose

Table 4. K_m-*values of galactinol : sucrose 6-galactosyltransferase*

Substrate	K_m value of	
	Synthesis reaction	Exchange reaction
	mM	mM
Galactinol	7.0	—
Raffinose	—	10
Sucrose[a]	1.0	2.0
Sucrose[b]	1.4	0.47

[a] In the presence of 0.02 M donor (*i. e.* galactinol and raffinose, respectively).
[b] In the presence of 0.2 mM donor.

Gal-αONp (an unphysiological substrate) and raffinose, *i.e.* in the exchange reaction, work to a significant extent (Table 5). Transfer from UDP-galactose to sucrose has been observed neither with the purified enzyme nor the crude extract [10].

Table 5. *Acceptor and donor specificity of galactinol:sucrose 6-galactosyltransferase (Sephadex fraction)*

In the acceptor experiment the incubation mixture contained in a total volume of 50 µl : 5 µmol Tris-HCl pH 7.2, 0.5 µmol acceptor, 0.039 µmol [^{14}C]galactinol (7 µCi/µmol) and 0.3 mg protein. After 4 h at 32 °C the reaction was stopped. In the donor experiment the incubation mixture contained 0.5 µmol donor, 0.02 µmol [^{14}C]sucrose (35 µCi/µmol), 5 µmol Tris-HCl and 0.1 mg protein. The incubation time was 1 h at 32 °C. Raffinose, stachyose, fructose, glucose, galactose, lactose, cellobiose, melibiose and glycerol do not act as acceptors

Donor	Acceptor	Reaction product	
		Raffinose	Galactose
		counts/min	
[^{14}C]Galactinol	Sucrose	11505	2495
[^{14}C]Galactinol	H_2O	0	10014
Galactinol	[^{14}C]Sucrose	24012	
UDP-Gal		0	
Gal-αONp		13210	
Raffinose		33405	
Stachyose		1005	
Melibiose		995	

DISCUSSION

The enzyme catalyzing the transfer of the galactosyl moiety from galactinol to sucrose has been isolated, purified and characterized. The results indicate that the enzyme is clearly different from any of the α-galactosidases described [5, 21—25]. Thus the hydrolyzing activity towards Gal-αONp, a typical substrate for α-galactosidases, separates from the raffinose-synthesizing enzyme during the purification. Furthermore the high substrate specificity as well as the efficiency of the transfer have to be pointed out, when the enzyme is compared with

α-galactosidases. It is proposed to call the enzyme galactinol : sucrose 6-galactosyl-transferase and to group it among the glycosyl transferases.

The exchange reaction is catalyzed by the same enzyme which is responsible for raffinose synthesis. This has also been expected in analogy to similar transfer reactions [7, 26, 27].

The enzyme activity of 7.9 nmol raffinose formed $\times h^{-1} \times g$ seeds^{-1} (Table 1) corresponds to an activity of 28.4 nmol $\times h^{-1} \times g$ seeds^{-1} at the physiological sucrose concentration of 10 mM. This rate is high enough to explain the synthesis rate *in vivo* for raffinose and for all the other higher homologues of the raffinose sugars during the ripening period. Thus the enzyme is able to synthesize 2.5 μmol raffinose, the amount actually present in 1 g of seeds, in less than 4 days.

The synthesis of the total amount of the other raffinose-type sugars (21.4 μmol/g seed) would take about one month, which corresponds reasonably well to the ripening period of the seeds.

In addition the results of the biosynthesis of raffinose and its higher homologues *in vitro* with respect to the function of galactinol are in agreement with the studies *in vivo* by Senser and Kandler [2, 29]. It seems without doubt now, that the biosynthesis of all the raffinose sugars proceeds *via* galactinol. The physiological meaning of the detour taken by the galactosyl moiety is not understood at present. Perhaps it has to be seen in relation to the observation that *myo*-inositol and galactinol inhibit α-galactosidases, enzymes responsible for the decomposition of raffinose sugars [9, 30].

Myo-inositol has been known as a growth factor for yeasts and many tissue cultures [31—33] for a long time. Since these cells do not contain sugars of the raffinose family the cofactor-like role, which *myo*-inositol plays in the biosynthesis of oligosaccharides, cannot explain its function as a growth factor. It seems likely, however, that *myo*-inositol is absolutely required in the form of phosphatidylinositols, which seem to be indispensable membrane components [34]. This is supported by the finding that transport mechanisms are impaired when cells lack *myo*-inositol [35—37].

We would like to thank Drs A. Böck and H. Kosakowski for helpful suggestions and advice.

REFERENCES

1. Cardini, C. E., Leloir, L. F. & Chiribiga, J. (1955) *J. Biol. Chem. 214*, 149—155.
2. Senser, M. & Kandler, O. (1967) *Z. Pflanzenphysiol. 57*, 376—388.
3. Tanner, W. & Kandler, O. (1966) *Plant. Physiol. 41*, 1540—1542.
4. French, D. (1954) *Adv. Carbohydr. Chem. 9*, 149—184.
5. Courtois, J. E. (1959) *Carbohydrate Chemistry of Substances of Biological Interest* (Wolfrom, M. L., ed.) pp. 140—150, Pergamon Press, London.
6. Jeffreys, K. (1962) *Bot. Stud. 15*, 1—96.
7. Tanner, W. & Kandler, O. (1968) *Eur. J. Biochem. 4*, 233—239.
8. Tanner, W., Lehle, L. & Kandler, O. (1967) *Biochem. Biophys. Res. Commun. 29*, 166—171.
9. Tanner, W. (1969) *Ann. N.Y. Acad. Sci. 165*, 726—742.
10. Lehle, L., Tanner, W. & Kandler, O. (1970) *Hoppe-Seyler's Z. Physiol. Chem. 351*, 1494—1498.
11. Bourne, E. J., Walter, M. W. & Pridham, J. B. (1965) *Biochem. J. 97*, 802—806.
12. Pridham, J. B. & Hassid, W. Z. (1965) *Plant. Physiol. 40*, 984—986.
13. Gomyo, T. & Nakamura, M. (1966) *Agric. Biol. Chem. 30*, 425—427.
14. Feydman, R. B. & Neufeld, E. F. (1963) *Biochem. Biophys. Res. Commun. 12*, 121—125.
15. Moreno, A. & Cardini, C. E. (1966) *Plant Physiol. 41*, 909—910.
16. Andrews, P. (1964) *Biochem. J. 91*, 222—233.
17. Maurer, H. R. (1968) *Disk Elektrophorese*, pp. 39—47, Walter de Gruyter and Co, Berlin.
18. Chrambach, A., Reisfeld, R. A., Wyckhoff, M. & Zaccari, J. (1967) *Anal. Biochem. 20*, 150—154.
19. Lowry, O. H., Rosebrough, N. J., Farr, A. L. & Randall, R. J. (1951) *J. Biol. Chem. 193*, 265—275.
20. Kandler, O. (1964) *Ber. Dtsch. Bot. Ges. 77*, 62—73.
21. Courtois, J. E. & Petek, F. (1960) *Methods Enzymol. 5*, 565—571.
22. Courtois, J. E. & Petek, F. (1957) *Bull. Soc. Chim. Biol. 39*, 715.
23. Dey, P. M. & Pridham, J. B. (1969) *Biochem. J. 115*, 47—54.
24. Malhotra, O. P. & Dey, P. M. (1967) *Biochem. J. 103*, 739—743.
25. Yu-Te Li & Shetlar, M. R. (1964) *Arch. Biochem. Biophys. 108*, 301—313.
26. Glaser, L. (1964) *Compr. Biochem. 15*, 93—137.
27. Nishizawa, K. & Hashimoto, J. (1970) *The Carbohydrates* (Pigman, W. & Horton, D., eds) vol. IIA, pp. 241—300, Academic Press, New York.
28. Tanner, W., Seifarth, H. & Kandler, O. (1968) *Z. Pflanzenphysiol. 58*, 369—377.
29. Senser, M. & Kandler, O. (1967) *Phytochemistry, 6*, 1533—1540.
30. Sharma, Ch. B. (1971) *Biochem. Biophys. Res. Commun. 43*, 672—679.
31. Burkholder, P. R., McVeigh, I. & Meyer, D. (1944) *J. Bacteriol. 48*, 385—391.
32. Eagle, H., Oyama, V. I., Levy, M. & Freeman, A. E. (1957) *J. Biol. Chem. 226*, 191—205.
33. Wolter, K. E. & Skoog, F. (1966) *Am. J. Bot. 53*, 203—209.
34. Jung, P., Tanner, W. & Wolter, K. (1972) *Phytochemistry, 11*, 1055—1059.
35. Lembach, K. & Chrambampous, F. C. (1967) *J. Biol. Chem. 242*, 2599—2605.
36. Chrambampous, F. C. (1969) *J. Biol. Chem. 244*, 1705—1710.
37. Scarborough, G. A. (1971) *Biochem. Biophys. Res. Commun. 43*, 968—975.

L. Lehle and W. Tanner, Fachbereich Biologie der Universität, D-8400 Regensburg, Universitätsstraße 31, Federal Republic of Germany

Planta (2002) 215: 839–846
DOI 10.1007/s00425-002-0804-7

Applicant's Copy 09/301,764

ORIGINAL ARTICLE

Thomas Peterbauer · Lukas Mach · Jan Mucha
Andreas Richter

Functional expression of a cDNA encoding pea (*Pisum sativum* L.) raffinose synthase, partial purification of the enzyme from maturing seeds, and steady-state kinetic analysis of raffinose synthesis

Received: 1 February 2002 / Accepted: 2 April 2002 / Published online: 25 June 2002

Abstract Raffinose (*O*-α-D-galactopyranosyl-(1→6)-*O*-α-D-glucopyranosyl-(1↔2)-*O*-β-D-fructofuranoside) is a widespread oligosaccharide in plant seeds and other tissues. Raffinose synthase (EC 2.4.1.82) is the key enzyme that channels sucrose into the raffinose oligosaccharide pathway. We here report on the isolation of a cDNA encoding for raffinose synthase from maturing pea (*Pisum sativum* L.) seeds. The coding region of the cDNA was expressed in *Spodoptera frugiperda* Sf21 insect cells. The recombinant enzyme, a protein of glycoside hydrolase family 36, displayed similar kinetic properties to raffinose synthase partially purified from maturing seeds by anion-exchange and size-exclusion chromatography. Apart from the natural galactosyl donor galactinol (*O*-α-D-galactopyranosyl-(1→1)-L-*myo*-inositol), *p*-nitrophenyl α-D-galactopyranoside, an artificial substrate, was utilized as a galactosyl donor. An equilibrium constant of 4.1 was determined for the galactosyl transfer reaction from galactinol to sucrose. Steady-state kinetic analysis suggested that raffinose synthase is a transglycosidase operating by a ping-pong reaction mechanism and may also act as a glycoside hydrolase. The enzyme was strongly inhibited by 1-deoxygalactonojirimycin, a potent inhibitor for α-galactosidases (EC 3.2.1.22). The physiological implications of these observations are discussed.

Keywords cDNA cloning · Enzyme kinetics · Galactinol · *Pisum* · Raffinose synthase · Seed

T. Peterbauer · A. Richter (✉)
Institute of Ecology,
University of Vienna, Althanstrasse 14,
1090 Vienna, Austria
E-mail: andreas.richter@univie.ac.at
Fax: +43-1-42779542

L. Mach · J. Mucha
Centre for Applied Genetics,
University of Agricultural Sciences Vienna,
Muthgasse 18, 1190 Vienna, Austria

Abbreviation PCR: polymerase chain reaction

Introduction

Raffinose (*O*-α-D-galactopyranosyl-(1→6)-*O*-α-D-glucopyranosyl-(1↔2)-*O*-β-D-fructofuranoside) and its higher homologue stachyose are major soluble carbohydrates in seeds, roots and tubers of many plant species (Avigad and Dey 1997). In the Lamiaceae, Cucurbitaceae, Oleaceae and other plant families, they are the predominant carbohydrates translocated in the phloem. Apart from carbon transport and storage, these oligosaccharides may function as protective agents during maturation drying of seeds (Horbowicz and Obendorf 1994) and during cold stress (Gilmour et al. 2000). The biosynthesis of raffinose was initially proposed to proceed by galactosyl transfer from UDP-galactose to sucrose (Bourne et al. 1965; Pridham and Hassid 1965; Imhoff 1973), but the direct nucleotide pathway has been disputed (Lehle and Tanner 1973; Bachmann et al. 1994). It is now generally accepted that raffinose is synthesized by the following reaction sequence:

$$\text{UDP} - \text{galactose} + \textit{myo}-\text{inositol} \rightleftharpoons \text{UDP} + \text{galactinol} \quad (1)$$

$$\text{Galactinol} + \text{sucrose} \rightleftharpoons \textit{myo}-\text{inositol} + \text{raffinose} \quad (2)$$

Reactions 1 and 2 are catalyzed by galactinol synthase (EC 2.4.1.123) and raffinose synthase (EC 2.4.1.82), respectively. Raffinose in turn is the substrate for stachyose synthase (EC 2.4.1.67), which adds a further galactose unit from galactinol:

$$\text{Galactinol} + \text{raffinose} \rightleftharpoons \textit{myo}-\text{inositol} + \text{stachyose} \quad (3)$$

Galactinol synthases and stachyose synthases have been characterized and corresponding cDNA sequences have been cloned from several sources (for review, see Peterbauer and Richter 2001). In contrast, raffinose synthase has attracted little attention, probably because it

is the most labile enzyme of the pathway. Since the pioneering work of Lehle and Tanner (1973), who purified the enzyme 400-fold from seeds of *Vicia faba*, raffinose synthase has only been characterized in a crude preparation from leaves of *Ajuga reptans* (Bachmann et al. 1994), although genes encoding for raffinose synthases have been reported in patents (Oosumi et al. 1998; Watanabe and Oeda 1998). At the amino acid level, these sequences show homology to stachyose synthases (Peterbauer et al. 1999, 2002) and seed imbibition proteins of unknown enzymatic function (Anderson and Kohorn 2001; Romo et al. 2001).

We have recently described the changes in the activity of raffinose synthase and other enzymes of the pathway during seed development of pea cultivars (Peterbauer et al. 2001). Here we describe the isolation and heterologous expression of a cDNA encoding a raffinose synthase from developing pea seeds. The kinetic properties of the recombinant enzyme are compared with those of a partially purified raffinose synthase preparation. We demonstrate that pea raffinose synthase is a transglycosidase with structural and biochemical similarities to α-galactosidases.

Materials and methods

Plant material and chemicals

Seeds of pea (*Pisum sativum* L. cv. Wunder von Kelvedon) were obtained from a local supplier (Austrosaat, Vienna, Austria). Plants were grown in commercial potting soil in a growth chamber at 22/18 °C day/night temperature and 50/80% relative humidity with a 16-h photoperiod. Seeds were harvested 20–30 days after flowering and were stored in liquid nitrogen.

Galactinol (*O*-α-D-galactopyranosyl-(1→1)-L-*myo*-inositol), D-ononitol (1D-4-*O*-methyl-*myo*-inositol), galactosyl ononitol (*O*-α-D-galactopyranosyl-(1→3)-4-*O*-methyl-D-*myo*-inositol) and galactopinitol A (*O*-α-D-galactopyranosyl-(1→2)-4-*O*-methyl-D-*chiro*-inositol) were available from previous studies (Wanek and Richter 1995; Richter et al. 1997; Peterbauer et al. 1998). Sucrose and D-galactose were obtained from Fluka (Vienna, Austria). Stachyose was from Merck (Vienna, Austria) and D-pinitol (1D-3-*O*-methyl-*chiro*-inositol) was from Aldrich (Vienna, Austria). Raffinose, *myo*-inositol, *p*-nitrophenyl α-D-galactopyranoside and 1-deoxygalactonojirimycin (1,5-dideoxy-1,5-imino-D-galactitol) were from Sigma (Vienna, Austria).

Reverse transcription-polymerase chain reaction (PCR) and rapid amplification of cDNA ends

Total RNA was extracted with an RNeasy Plant Mini kit and poly(A)$^+$ RNA was isolated with an Oligotex mRNA kit (Qiagen, Hilden, Germany). First-strand cDNA was synthesized from poly(A)$^+$ RNA using AMV reverse transcriptase and an oligo-(dT) primer. Amplification by PCR was performed with HotStar *Taq* DNA polymerase (Qiagen) and degenerate primers. The sense xprimer 5′-TT(T/C)GGITGGTG(C/T)ACITGGGA(T/C)GC-3′ (where I denotes inosine) was based on the amino acid sequence motif FGWCTWDA. The antisense primer 5′-CCAIC-CI(G/C)CI CC(C/T)TG(A/G)CA(G/A)TT(G/A)AA-3′ was based on the motif FNCQG(A/G)GW. PCR products were isolated from agarose gels, cloned into the pCR2.1-TOPO vector (Invitrogen, Lofer, Austria) and sequenced using an ABI Prism BigDye Terminator Cycle Sequencing Mix and an ABI Prism 310 sequencer (Applied Biosystems, Vienna, Austria). RNA ligase-mediated rapid amplification of the missing cDNA ends was performed with the GeneRacer system (Invitrogen) as suggested by the manufacturer. The 5′-end was amplified with the gene-specific primer 5′-CGGTT CATTCCATCTCGCTCTGTAA-3′. The 3′-end was amplified with the gene-specific primer 5′-TTGTTTTTGCCCGA CGGTTCT ATCTT-3′ followed by nested PCR with the gene-specific primer 5′-GTCAACATTACGCACTCCCTACACGA-3′. PCR products were cloned and sequenced as described above. The assembled cDNA sequence was deposited in the EBI database under the accession number AJ426475.

Expression of raffinose synthase in baculovirus-infected insect cells

The open reading frame encoded by the putative raffinose synthase cDNA was amplified from reverse-transcribed poly(A)$^+$ RNA using *Pfu* DNA polymerase (Promega, Mannheim, Germany) and the primers 5′-CTGCAGGCACCACCAAGCATAAC-3′ (sense) and 5′-GGTACCCATGAGGGATCAAAATAAAAAC-3′ (antisense). The PCR product was cloned into pCR2.1-TOPO and sequenced. Restriction sites for *Pst*I and a *Kpn*I (underlined) were included in the PCR primers for subsequent subcloning of the fragment, in frame, into the baculovirus expression vector pVTBacHis-1 (Sarkar et al. 1998). Co-transfection of the expression vector with linearized viral DNA and amplification of recombinant baculovirus in *Spodoptera frugiperda* insect cells was performed as previously described (Mucha et al. 2001). Infected Sf21 insect cells expressing the recombinant protein under control of the polyhedrin promoter were lysed in 50 mM Na-phosphate (pH 7.0), 1 mM DTT, containing a set of protease inhibitors (Complete Protease Inhibitor Cocktail; Roche, Vienna, Austria). Cell lysates were desalted by repeated ultrafiltration in 50 mM Na-phosphate (pH 7.0), 1 mM DTT, using Centricon Plus-20 ultrafiltration units (Millipore, Vienna, Austria), and assayed for activity. Western blot analysis of crude cell culture supernatants with monoclonal antibodies against an enterokinase site provided by the expression vector was performed as previously described (Mucha et al. 2001).

Partial purification of raffinose synthase from pea seeds

All steps were carried out at 4 °C unless otherwise stated. Maturing seeds (95 g) were immersed in liquid nitrogen, ground to a fine powder and suspended in 150 ml of 50 mM Hepes-NaOH (pH 7.0), 20 mM $MgCl_2$, 2.5 mM EGTA, 0.5 mM DTT, 1% polyvinylpolypyrrolidone. The extract was further homogenized with a Polytron tissue homogenizer, filtered through one layer of fine-mesh nylon and centrifuged at 26,000 *g* for 30 min. The supernatant was adjusted with stirring to 2 mg ml^{-1} protamine sulfate by dropwise addition of a 10% (w/v) protamine sulfate solution (to prevent clogging of columns in subsequent chromatography) in 50 mM Hepes-NaOH (pH 7.0). Precipitated material (nucleic acids, some storage proteins and other contaminants) was removed by centrifugation at 26,000 *g* for 20 min. The cleared supernatant was subjected to fractionation with solid ammonium sulfate. Proteins precipitating between 35 and 55% saturation were collected by centrifugation at 26,000 *g* for 20 min, dissolved in 20 mM bis-Tris propane-HCl (pH 6.8), 0.5 mM DTT, and dialyzed overnight against this buffer. The sample was loaded onto an anion-exchange column (90 ml, 2.5 cm i.d.) of Macro-Prep High Q (Bio-Rad) maintained at 12 °C. Bound protein was eluted at a flow rate of 5 ml min^{-1} with a linear gradient (750 ml) of 0–250 mM NaCl in 20 mM bis-Tris propane-HCl (pH 6.8), 0.5 mM DTT. Active fractions were pooled, concentrated by ultrafiltration, and applied at a flow rate of 0.5 ml min^{-1} to a Superdex 200 HR 10/30 size-exclusion chromatography column (Amersham Pharmacia Biotech, Vienna, Austria) equilibrated with 20 mM Na-phosphate (pH 7.0), 1 mM DTT, 150 mM NaCl. Active fractions were pooled, concentrated by ultrafiltration, and stored in aliquots in liquid nitrogen.

Enzyme and protein assay

Raffinose synthase activity was routinely determined at 30 °C in reaction mixtures (20 µl) containing 50 mM Na-phosphate (pH 7.0), 1 mM DTT, 10 mM galactinol and 20 mM sucrose. Enzyme samples were diluted and incubation times were adjusted to allow transformation of not more than 10% of the substrates into products. Reactions were stopped by boiling for 5 min. Reaction mixtures were diluted to 0.5 ml and centrifuged for 5 min at 12,000 *g*. Formation of raffinose and galactose (arising by hydrolysis of galactinol) was determined by HPLC with pulsed amperometric detection using a Carbopac PA-10 column (250 mm long, 2 mm i.d.; Dionex, Vienna, Austria) as previously described (Peterbauer et al. 2002). Formation of *myo*-inositol and galactinol (a product in the reverse reaction) was determined by HPLC using a Carbopac MA-1 column (Dionex) thermostatted at 25 °C. Sugars were eluted with 150 mM NaOH at a flow rate of 0.4 ml min^{-1}. After each run, the column was washed with 1 M NaOH to elute raffinose. Synthesis of galactosyl ononitol and galactopinitol A by galactosyl transfer from galactinol to D-ononitol and D-pinitol, respectively, was determined by capillary gas chromatography as previously described (Peterbauer et al. 1998; Hoch et al. 1999). The concentration of soluble protein was determined by the dye-binding procedure (Bradford 1976) using BSA as a standard.

Steady-state kinetic analysis and equilibrium

For the determination of kinetic constants of raffinose synthesis, samples were incubated with varying concentrations of the first substrate at several fixed concentrations of the second substrate. Data were fitted to the initial rate equation for a ping-pong Bi Bi mechanism:

$$v = \frac{V_{max}[A][B]}{K_{mB}[A] + K_{mA}[B] + [A][B]} \quad (4)$$

where *v* is the initial velocity, V_{max} is the maximum velocity, [A] and [B] are the concentrations of the substrates, and K_{mA} and K_{mB} are Michaelis constants for A and B, respectively. Inhibition patterns were determined graphically by replots of slopes and intercepts of primary double-reciprocal plots. Inhibition constants were estimated by fitting the untransformed data to Eq. 5 or 6, corresponding to linear competitive or mixed (non-competitive) inhibition, respectively, or to Eq. 7, corresponding to an hyperbolic uncompetitive inhibition pattern (Cleland 1963):

$$v = \frac{V_{max}[S]}{K_m(1 + [I]/K_{ic}) + [S]} \quad (5)$$

$$v = \frac{V_{max}[S]}{K_m(1 + [I]/K_{ic}) + [S](1 + [I]/K_{iu})} \quad (6)$$

$$v = \frac{V_{max}[S]}{K_m + [S](1 + [I]/K_{iu})/(1 + [I]/K'_{iu})} \quad (7)$$

where [S] is the concentration of the variable substrate, [I] is the concentration of the inhibitor, K_{ic} is a competitive inhibition constant, and K_{iu} and K'_{iu} are uncompetitive inhibition constants, respectively.

To estimate K_{eq} of raffinose synthesis, the partially purified enzyme (230 pkat) was incubated in a final volume of 75 µl with 40 mM galactinol and sucrose, 40 mM raffinose and *myo*-inositol, or with all four substrates (20 mM each), respectively. At intervals, aliquots were removed and analyzed by HPLC. Kinetic and thermodynamic constants are given as means ± SE.

Results

cDNA cloning and analysis of pea raffinose synthase

To isolate a cDNA encoding for raffinose synthase by reverse transcription-PCR, degenerate oligonucleotide primers were designed based on amino acid motifs conserved among *Cucumis sativus* raffinose synthase, stachyose synthases and related sequences (Peterbauer et al. 1999). To distinguish between raffinose synthase and stachyose synthase, the primers were chosen to encompass a block of about 80 amino acids, which is exclusively present in stachyose synthases. Two PCR products of about 1.2 and 1.4 kbp, respectively, were obtained from mRNA isolated from maturing seeds. Upon sequence analysis, the longer fragment revealed 100% identity with pea stachyose synthase (Peterbauer et al. 2002). The 5′- and the 3′-end of the 1.2-kbp fragment were extended by RNA ligase-mediated rapid amplification. The composed sequence of 2,652 nucleotides contains an open reading frame of 2,394 nucleotides encoding for a polypeptide of 798 amino acids with a calculated molecular mass of 88.7 kDa. All other methionine codons were found to be in-frame with the putative start codon.

A pattern search using the BLOCKS database (Henikoff et al. 2000) revealed the presence of the glycoside hydrolase superfamily GH-D signature. According to the sequence-based classification of Henrissat (Henrissat and Davies 2000), this superfamily is formed by α-galactosidases from glycoside hydrolase families 27 and 36 (Dagnall et al. 1995). An alignment of one of the characteristic motifs with members of the GH-D superfamily is shown in Fig. 1. It contains a conserved aspartic acid residue, which acts as a catalytic nucleophile to generate a covalent glycosyl-enzyme intermediate in α-galactosidases of family 27 (Hart et al. 2000; Ly et al. 2000). High overall sequence homology with seed imbibition proteins (SIPs) and stachyose synthases (not shown) places the pea protein into the related glycoside hydrolase family 36 (for database entries, see http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html).

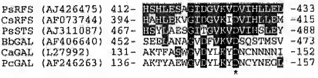


Fig. 1. Partial alignment of pea (*Pisum sativum*) raffinose synthase with selected members of the glycoside hydrolase superfamily GH-D. *PsRFS*, *P. sativum* raffinose synthase; *CsRFS*, *Cucumis sativus* raffinose synthase; *PsSTS*, *P. sativum* stachyose synthase; *BbGal*, *Bifidobacterium breve* α-galactosidase; *CaGAL*, *Coffea arabica* α-galactosidase (preprotein); *PcGAL*, *Phanerochaete chrysosporium* α-galactosidase (preprotein). EBI/GenBank accession numbers are shown in parentheses. The catalytic aspartic acid residue in *C. arabica* and *P. chrysosporium* α-galactosidase is marked by an *asterisk*. The alignment was generated using CLUSTAL W (Thompson et al. 1994)

Expression of raffinose synthase in insect cells

The coding region of the cDNA was engineered into a baculovirus expression vector and inserted into the baculovirus genome by homologous recombination. Desalted lysates of insect cells infected with recombinant virus displayed raffinose synthase activity (0.82 pkat mg^{-1} soluble protein), while no synthesis of raffinose was detected in lysates of uninfected control cells. The recombinant protein, which was fused to a leader sequence containing a honeybee melittin signal peptide for secretion, a hexahistidine tag and an enterokinase cleavage site, was also detected in the culture medium of infected insect cells by Western blot analysis with monoclonal antibodies against the enterokinase site (Fig. 2). However, several attempts to purify the protein from the medium by chromatography on iminodiacetic acid-Sepharose charged with Ni^{2+} failed. Only traces of activity were recovered, probably because Ni^{2+} destabilizes the enzyme. When insect cell lysates were incubated for 1 h with 1 mM $NiCl_2$ in solution, activity decreased to 30.2% of untreated controls. The steady-state kinetics of raffinose synthesis were therefore analyzed using lysates of insect cells. A double-reciprocal plot revealed a set of apparently parallel lines (Fig. 3). This pattern is characteristic of a ping-pong Bi Bi mechanism, in which a glycosyl–enzyme intermediate of any kind is formed and the first product dissociates before the second substrate is bound. Kinetic constants were estimated as described in *Materials and methods* (Table 1). Due to endogenous α-galactosidase activity in insect cells, no attempts were made to characterize hydrolytic activity of the recombinant protein (see below for further analysis).

Formation of raffinose was also observed when galactinol was replaced by galactosyl ononitol, a methylated derivative of galactinol, or by the artificial substrate *p*-nitrophenyl α-D-galactopyranoside (Table 2). On the other hand, sucrose could be replaced by D-ononitol and D-pinitol, yielding galactosyl ononitol and a galactopinitol, respectively. Raffinose itself was not utilized as an acceptor.

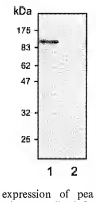


Fig. 2. Heterologous expression of pea raffinose synthase in *Spodoptera frugiperda* insect cells infected with recombinant baculovirus. Crude cell culture supernatants were subjected to Western blot analysis with monoclonal mouse antibody raised against the enterokinase recognition sequence fused to the recombinant protein. Bound antibody was visualized with anti-mouse antibodies conjugated to horseradish peroxidase and chemiluminescence detection. *Lane 1* Supernatant from infected insect cells, *lane 2* supernatant from uninfected control cells

Partial purification and characterization of raffinose synthase from maturing seeds

After sample clean-up by treatment with protamine sulfate and ammonium sulfate fractionation, raffinose synthase was partially purified from a pea seed extract by anion- and size-exclusion chromatography. The final preparation had a specific activity of 75.4 pkat mg^{-1} protein. The K_m values for galactinol and sucrose were experimentally indistinguishable from those determined for the recombinant raffinose synthase (Table 1). Relative activities towards other donor and acceptor substrates were also similar (Table 2). *myo*-Inositol acted as a linear mixed product inhibitor with respect to galactinol, while linear competitive inhibition was observed with respect to sucrose as the varied substrate (Table 3). Partially purified raffinose synthase was strongly inhib-

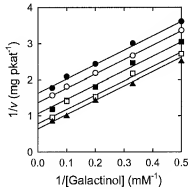


Fig. 3. Initial velocity pattern of raffinose synthesis catalyzed by recombinant pea raffinose synthase in lysates of insect cells. The concentrations of sucrose were 10 mM (*filled circles*), 13 mM (*open circles*), 20 mM (*filled squares*), 40 mM (*open squares*) or 80 mM (*triangles*). Lines represent the best fit to Eq. 4

Table 1. Kinetic parameters for the synthesis of raffinose and hydrolysis of galactinol catalyzed by insect cell lysates expressing the recombinant enzyme and by raffinose synthase partially purified from pea (*Pisum sativum*) seeds, respectively. Kinetic constants were estimated by non-linear regression as described in *Materials and methods*

Reaction	V_{max} (pkat mg^{-1})	K_m Galactinol (mM)	K_m Sucrose (mM)
Insect cell lysates			
Raffinose synthesis	2.0 ± 0.1	7.9 ± 0.7	22.6 ± 2.2
Partially purified enzyme			
Raffinose synthesis	199.2 ± 7.8	7.3 ± 0.5	22.9 ± 0.6
Galactinol hydrolysis	27.3 ± 0.3	1.0 ± 0.1	

Table 2. Substrate specificity of recombinant raffinose synthase determined in lysates of insect cells and of a partially purified raffinose synthase preparation from maturing pea seeds. Reaction mixtures contained galactosyl donors and acceptors at concentrations of 10 and 20 mM, respectively. *n.d.* Not detected

Donor	Acceptor	Relative activity (%)	
		Insect cell lysates	Partially purified raffinose synthase
Galactinol	Sucrose	100.0	100.0
Galactinol	Raffinose	n.d.	n.d.
Galactinol	D-Ononitol	86.6	105.1
Galactinol	D-Pinitol	42.3	32.6
Galactosyl ononitol	Sucrose	67.1	68.3
p-Nitrophenyl α-D-galactopyranoside	Sucrose	33.5	39.4

ited by 1-deoxygalactonojirimycin (Fig. 4). Inhibition was found to be competitive with respect to galactinol, with an apparent K_{ic} of 189 ± 14 nM. Stachyose (50 mM) had no effect on raffinose synthase activity.

The enzyme exhibited an optimum at pH 7.0 (Fig. 5). At the optimum of pea α-galactosidase (pH 4.5; Peterbauer et al. 2001), very little hydrolysis of galactinol was detected, suggesting that the preparation was essentially free of acidic α-galactosidases. However, some hydrolysis occurred around pH 7.0. When the rate of release of galactose at pH 7.0 was considered as the reaction rate, sucrose acted as linear mixed inhibitor (Fig. 6a). Inhibition constants are compiled in Table 3. When the rate of release of *myo*-inositol at several fixed levels of sucrose was plotted, parallel lines were obtained (Fig. 6b). A replot of intercepts of the primary double-reciprocal plot as a function of the sucrose concentration revealed hyperbolic activation. These patterns of inhibition of the hydrolytic activity together with those of the transfer reaction are unique for a ping-pong mechanism where an unstable glycosyl–enzyme complex is formed[1] (Cleland 1963; 1970). In other words, the intermediary glycosyl–enzyme complex either reacts with sucrose to give raffinose, or hydrolyzes to give galactose and free enzyme. The observation that sucrose acted as an activator (and not as an inhibitor) with respect to the rate of *myo*-inositol formation indicates that the release of raffinose is faster than hydrolysis of the glycosyl–enzyme complex.

Equilibrium of raffinose synthesis

To estimate K_{eq} for raffinose synthesis, the enzyme was incubated with varying concentrations of substrates and products (Fig. 7). Starting from either side of the reaction, mass action ratios approached similar values. From the substrate and product concentrations obtained after 5 h, a mean mass action ratio of 4.1 ± 0.6 was calculated for the synthesis reaction. Since hydrolysis of substrates was slow compared with the transfer reactions, it is reasonable to assume that K_{eq} is close to 4.

[1]This special ping-pong mechanism predicts convergent rather than parallel lines in double-reciprocal plots of the rates of raffinose formation. Within substrate ranges used here, however, hydrolysis of galactinol was too slow, as compared with the transfer reaction, to allow detection of convergence experimentally (see Fig. 3).

Discussion

A cDNA encoding for raffinose synthase was isolated from maturing pea seeds and functionally expressed in insect cells. The kinetic properties of the recombinant protein were similar to those of partially purified raffinose synthase (Table 1), providing good evidence that the cloned cDNA corresponds to the enzyme expressed in developing seeds. The synthesis reaction was reversible with a K_{eq} of about 4 (Fig. 7). This value is very similar to the equilibrium of stachyose synthesis (Tanner and Kandler 1968). Like the corresponding enzyme from *Vicia faba* (Lehle and Tanner 1973), pea raffinose synthase displayed an optimum at pH 7.0 (Fig. 5). Steady-state kinetics (Fig. 3) and product inhibition by *myo*-inositol (Table 3) suggested that the synthesis of raffinose proceeds by a ping-pong mechanism. This mechanism explains isotopic exchange between raffinose and labelled sucrose, which has been shown to be associated with raffinose synthase activity (Lehle and Tanner 1973; Castillo et al. 1990). Remarkably, pea raffinose synthase was able to utilize D-ononitol and D-pinitol as acceptors (Table 2). Galactosyl transfer to these methylated inositols is similar to an exchange reaction between galactinol and *myo*-inositol, which is also expected to occur in a ping-pong mechanism. However, it has so far been believed that these reactions are exclusively catalyzed by stachyose synthases (Peterbauer and Richter 2001). Our results indicate that the ability to utilize various inositol derivatives is a more common feature of enzymes of the raffinose oligosaccharide pathway.

Direct evidence for hydrolytic activity of raffinose synthase towards galactinol could not be provided, because we were unable to purify recombinant protein. However, indirect support came from steady-state kinetic analysis of galactinol hydrolysis catalyzed by the partially purified protein (Fig. 6). A dual role of raffinose synthase as a transglycosidase with some hydrolytic activity is in line with a classification of raffinose synthase as a glycoside hydrolase of family 36. Members of a family are likely to share structural topology of the active site and a common catalytic mechanism, although overall sequence homologies may be weak (Henrissat and Davies 2000). Indeed, all α-galactosidases so far studied, as well as stachyose synthases of family 36, have been shown to operate by a ping-pong mechanism (Peterbauer and Richter 1998; Brumer et al. 1999; Van Laere et al. 1999; Peterbauer et al. 2002). It is interesting to note that the amino acid residue, which forms the

Table 3. Inhibition patterns and inhibition constants for partially purified raffinose synthase from maturing pea seeds. Inhibition constants were determined by fitting the data to the corresponding initial rate equations as described in *Materials and methods*. *C* Linear competitive inhibition, *M* linear mixed (non-competitive) inhibition, *hU* hyperbolic uncompetitive inhibition

Variable substrate	Product	Inhibitor	Pattern	K_{ic} (mM)	K_{iu} (mM)	K'_{iu} (mM)
Sucrose	Raffinose	*myo*-Inositol	C	10.1 ± 0.9[a]		
Galactinol	Raffinose	*myo*-Inositol	M	22.3 ± 4.1[b]	23.2 ± 2.6[b]	
Galactinol	Galactose	Sucrose	M	3.7 ± 0.5	22.8 ± 2.8	
Galactinol	*myo*-inositol	Sucrose	hU		18.7 ± 1.3	2.8 ± 0.2

[a]Apparent inhibition constant determined in the presence of 10 mM galactinol
[b]Apparent inhibition constant determined in the presence of 20 mM sucrose

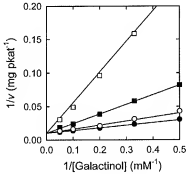


Fig. 4. Inhibition of partially purified raffinose synthase from maturing pea seeds by 1-deoxygalactonojirimycin. Assays contained 2–20 mM galactinol and 20 mM sucrose. The concentrations of 1-deoxygalactonojirimycin were 0.0 μM (*filled circles*), 0.1 μM (*open circles*), 0.5 μM (*filled squares*) or 2.0 μM (*open squares*). Lines represent the best fit to Eq. 5

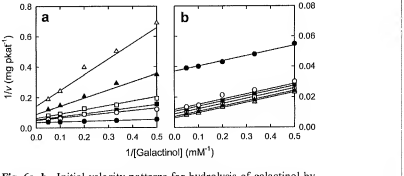


Fig. 6a, b. Initial velocity patterns for hydrolysis of galactinol by partially purified raffinose synthase from maturing pea seeds in the presence of several fixed concentrations of sucrose. **a** Release of galactose. Lines represent the best fit to Eq. 6. **b** Formation of *myo*-inositol. Lines represent the best fit to Eq. 7. The concentrations of sucrose were 0 mM (*filled circles*), 10 mM (*open circles*), 13 mM (*filled squares*), 20 mM (*open squares*), 40 mM (*filled triangles*) or 80 mM (*open triangles*)

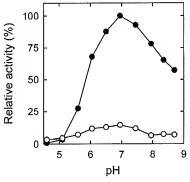


Fig. 5. Effect of pH on the formation of raffinose (*filled circles*) and release of galactose (*open circles*) in reaction mixtures containing partially purified raffinose synthase from maturing pea seeds, 10 mM galactinol and 20 mM sucrose in McIlvaine buffer (0.2 M Na_2HPO_4 adjusted to various pH values with 0.1 M citric acid). Data were adjusted relative to the maximum activity measured

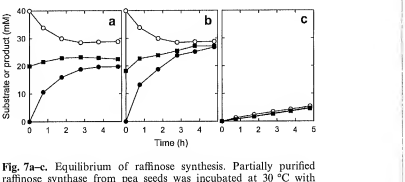


Fig. 7a–c. Equilibrium of raffinose synthesis. Partially purified raffinose synthase from pea seeds was incubated at 30 °C with 40 mM galactinol and 40 mM sucrose (*filled circles*), 20 mM of each galactinol, *myo*-inositol, sucrose and raffinose (*squares*), or 40 mM *myo*-inositol and 40 mM raffinose (*open circles*), respectively. Aliquots were removed at the times indicated and analyzed by HPLC. Changes in the levels of substrates and products are shown for raffinose (**a**), *myo*-inositol (**b**) and galactose (**c**)

covalent intermediate in α-galactosidases of the related glycoside hydrolase family 27, is conserved among members of family 36 (Fig. 1).

Two other biochemical properties of the enzyme further support a structural similarity of the active site of raffinose synthase and α-galactosidases. Like the corresponding enzyme from *Vicia faba* (Lehle and Tanner 1973), pea raffinose synthase utilized *p*-nitrophenyl α-D-galactopyranoside, an artificial substrate with high affinity towards α-galactosidases, as a galactosyl donor (Table 2). The enzyme was also strongly inhibited by 1-deoxygalactonojirimycin (Fig. 4), a potent competitive inhibitor for α-galactosidases (Asano et al. 2000; Martin et al. 2001). These findings may be of

more general relevance. A large number of α-galactosidases with acidic or alkaline pH optima have been identified using nitrophenyl derivatives as substrates (for reviews, see Dey 1985; Peterbauer and Richter 2001). With the exception of a few α-galactosidases of bacterial origin (van den Broek et al. 1999; Van Laere et al. 1999), none of these enzymes has been rigorously tested for transglycosidase activities with natural substrates, because it is tacitly assumed that proteins, which act on nitrophenyl glycosides in vitro, function as hydrolases in vivo. Our results suggest that this assumption may be misleading. Raffinose synthase, for example, would be recognized as neutral α-galactosidase if assayed only with *p*-nitrophenyl galactopyranoside or galactinol. However, its hydrolytic activity is probably of little physiological significance, because it is strongly inhibited by sucrose (Fig. 6a), which is usually present in high concentration in plant tissues.

Inhibition of raffinose synthase by 1-deoxygalactonojirimycin is of particular interest, because this inhibitor has been successfully used to modulate activity of α-galactosidase in human lymphoblasts (Asano et al. 2000). Hence, it could be possible to use 1-deoxygalactonojirimycin or related α-galactosidase inhibitors to experimentally manipulate the content of raffinose oligosaccharides in vivo. A more detailed characterization of enzymes involved in raffinose oligosaccharide synthesis with respect to inhibition by 1-deoxygalactonojirimycin is currently in progress.

Acknowledgements We thank Barbara Svoboda for excellent technical support. This work was supported by the Austrian Science Fund (FWF) Grant P13955-BIO (to A.R.).

References

Anderson CM, Kohorn BD (2001) Inactivation of *Arabidopsis SIP1* leads to reduced levels of sugars and drought tolerance. J Plant Physiol 158:1215–1219

Asano N, Ishii S, Kizu H, Ikeda K, Yasuda K, Kato A, Martin OR, Fan JQ (2000) In vitro inhibition and intracellular enhancement of lysosomal α-galactosidase A activity in Fabry lymphoblasts by 1-deoxygalactonojirimycin and its derivatives. Eur J Biochem 267:4179–4186

Avigad G, Dey PM (1997) Carbohydrate metabolism: storage carbohydrates. In: Dey PM, Harbourne J (eds) Plant biochemistry. Academic Press, San Diego, pp 143–204

Bachmann M, Matile P, Keller F (1994) Metabolism of the raffinose family oligosaccharides in leaves of *Ajuga reptans* L. Cold acclimation, translocation, and sink to source transition: discovery of a chain elongation enzyme. Plant Physiol 105:1335–1345

Bourne EJ, Walter MW, Pridham JB (1965) The biosynthesis of raffinose. Biochem J 97:802–806

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye-binding. Anal Biochem 72:248–254

Brumer H, Sims PFG, Sinnot ML (1999) Lignocellulose degradation by *Phanerochaete chrysosporium*: purification and characterization of the main α-galactosidase. Biochem J 339:43–53

Castillo EM, de Lumen BO, Reyes PS, de Lumen HZ (1990) Raffinose synthase and galactinol synthase in developing seeds and leaves of legumes. J Agric Food Chem 38:351–355

Cleland WW (1963) The kinetics of enzyme-catalyzed reactions with two or more substrates or products II. Inhibition: nomenclature and theory. Biochim Biophys Acta 67:173–187

Cleland WW (1970) Steady state kinetics. In: Boyer PD (ed) The enzymes, vol II. Kinetics and mechanism. Academic Press, New York, pp 1–66

Dagnall BH, Paulsen IT, Saier MH (1995) The DAG family of glycosyl hydrolases combines two previously identified protein families. Biochem J 311:349–350

Dey PM (1985) D-Galactose-containing oligosaccharides. In: Dey PM, Dixon R (eds) Biochemistry of storage carbohydrates in green plants. Academic Press, New York, pp 53–129

Gilmour SJ, Sebolt AM, Salazar MP, Everard JD, Thomashow MF (2000) Overexpression of the *Arabidopsis CBF3* transcriptional activator mimics multiple biochemical changes associated with cold acclimation. Plant Physiol 124:1854–1865

Hart DO, He SM, Chany CJ, Withers SG, Sims PFG, Sinnott ML, Brumer H (2000) Identification of Asp-130 as the catalytic nucleophile in the main α-galactosidase from *Phanerochaete chrysosporium*, a family 27 glycosyl hydrolase. Biochemistry 39:9826–9836

Henikoff JG, Greene EA, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the Blocks Database servers. Nucleic Acids Res 28:228–230

Henrissat B, Davies GJ (2000) Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. Plant Physiol 124:1515–1519

Hoch G, Peterbauer T, Richter A (1999) Purification and characterization of stachyose synthase from lentil (*Lens culinaris*) seeds: galactopinitol and stachyose synthesis. Arch Biochem Biophys 366:75–81

Horbowicz M, Obendorf RL (1994) Seed desiccation tolerance and storability: dependence on flatulence-producing oligosaccharides and cyclitols – review and survey. Seed Sci Res 4:385–405

Imhoff V (1973) Synthesis of galactosides by chloroplasts isolated from pea leaves. Hoppe-Seyler's Z Physiol Chem 354:1550–1554

Lehle L, Tanner W (1973) The function of *myo*-inositol in the biosynthesis of raffinose. Purification and characterization of galactinol:sucrose 6-galactosyltransferase from *Vicia faba* seeds. Eur J Biochem 38:103–110

Ly HD, Howard S, Shum K, He S, Zhu A, Withers SG (2000) The synthesis, testing and use of 5-fluoro-α-D-galactosyl fluoride to trap an intermediate on green coffee bean α-galactosidase and identify the catalytic nucleophile. Carbohydr Res 329:539–547

Martin OR, Saavedra OM, Xie F, Liu L, Picasso S, Vogel P, Kizu H, Asano N (2001)α- and β-homogalactonojirimycins (α- and β-homogalactostatins): synthesis and further biological evaluation. Bioorg Med Chem 9:1269–1278

Mucha J, Svoboda B, Fröhwein U, Strasser R, Mischinger M, Schwihla H, Altmann F, Hane W, Schachter H, Glössl J, Mach L (2001) Tissues of the clawed frog *Xenopus laevis* contain two closely related forms of UDP-GlcNAc:α3-D-mannoside β-1,2-*N*-acetylglucosaminyltransferase I. Glycobiology 11:769–778

Oosumi C, Nozaki J, Kida T (1998) Raffinose synthetase gene, process for producing raffinose, and transformed plant. International Patent Publication WO98/49273, PCT/JP97/03879, November 5, 1998

Peterbauer T, Richter A (1998) Galactosylononitol and stachyose synthesis in seeds of adzuki bean. Purification and characterization of stachyose synthase. Plant Physiol 117:165–172

Peterbauer T, Richter A (2001) Biochemistry and physiology of raffinose family oligosaccharides and galactosyl cyclitols in seeds. Seed Sci Res 11:185–197

Peterbauer T, Puschenreiter M, Richter A (1998) Metabolism of galactosylononitol in seeds of *Vigna umbellata*. Plant Cell Physiol 39:334–341

Peterbauer T, Mucha J, Mayer U, Popp M, Glössl J, Richter A (1999) Stachyose synthesis in seeds of adzuki bean (*Vigna angularis*): molecular cloning and functional expression of stachyose synthase. Plant J 20:509–518

Peterbauer T, Lahuta LB, Blöchl A, Mucha J, Jones DA, Hedley CL, Górecki RJ, Richter A (2001) Analysis of the raffinose family oligosaccharide pathway in pea seeds with contrasting carbohydrate composition. Plant Physiol 127:1764–1772

Peterbauer T, Mucha J, Mach L, Richter A (2002) Chain-elongation of raffinose in pea seeds. Isolation, characterization, and molecular cloning of a multifunctional enzyme catalyzing the synthesis of stachyose and verbascose. J Biol Chem 277:194–200

Pridham JB, Hassid WZ (1965) Biosynthesis of raffinose. Plant Physiol 40:984–986

Richter A, Peterbauer T, Brereton I (1997) Structure of galactosylononitol. J Nat Prod 60:749–751

Romo S, Labrador E, Dopico B (2001) Water stress-regulated gene expression in *Cicer arietinum* seedlings and plants. Plant Physiol Biochem 39:1017–1026

Sarkar M, Pagny S, Unligil U, Joziasse D, Mucha J, Glössl J, Schachter H (1998) Removal of 106 amino acids from the N-terminus of UDP-GlcNAc:α-D-mannoside β-1,2-*N*-acetylglucosaminyltransferase I does not inactivate the enzyme. Glycoconjugate J 15:193–197

Tanner W, Kandler O (1968) *myo*-Inositol, a cofactor in the biosynthesis of stachyose. Eur J Biochem 4:233–239

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

van den Broek LAM, Ton J, Verdoes JC, Van Laere KMJ, Voragen AGJ, Beldman G (1999) Synthesis of α-galacto-oligosaccharides by a cloned α-galactosidase from *Bifidobacterium adolescentis*. Biotechnol Lett 21:441–445

Van Laere KMJ, Hartemink R, Beldman G, Pitson S, Dijkema C, Schols HA, Voragen AGJ (1999) Transglycosidase activity of *Bifidobacterium adolescentis* DSM 20083 α-galactosidase. Appl Microbiol Biotechnol 52:681–688

Wanek W, Richter A (1995) Purification and characterization of *myo*-inositol 6-*O*-methyltransferase from *Vigna umbellata* Ohwi and Ohashi. Planta 197:1–8

Watanabe E, Oeda K (1998) Raffinose synthetase genes and use thereof. European Patent Application EP0849359, June 24, 1998

EXHIBIT 4

```
SC-05    1:MAPPSVIKSDAAVNGIDLSGKPLFRLEGSDLLANGHVVLTDVPVNVTVTASPYLADKDGEPVDASAGSFIGFNLDGEPRSRH 82
SC-07     :----------------------------------------------------------------------------------

                                  ->111
SC-05   83:VASIGKLRDIRFMSIFRFKVWWTTHWVGSKGSDIENETQIIFLLNSGSGEPYVLILPLLEGSFRSSFQPGEDDDVAVCVESG 164
SC-07     :----------------------------------------------------------------------------------

                                                 213<-
SC-05  165:STQVTQSDEQVVYYVHAGDDPFKLVKDAMKVVRVEMNTEELLEEKTPPGIVDKFGWCTWDAFYLTVNPDGVHKGVKCLVDGGCPPGLV 252
SC-07    1:-----------------------------------------LEEKTPPGIVDKFGWCTWDAFYLTVNPDGVHKGVKCLVDGGCPPGLV 47
                                                 ***********************************************

          ->260          275<-          ->293                         325<-
SC-05  253:LIDDGWQSIGHDSDGIDVEGMSCTVAGEQMPCRLLKFQENFKERDYVSPKDKNEVMKAFVRDLKEEFSTVDYIYVWHALCGYW 336
SC-07   48:LIDDGWQSIGHDSDGIDVEGMSCTVAGEQMPCRLPKFQENFKFRDYVSPKDKNEVMKAFVRDLKEEFSTVDYIYVWHALCGYW 131
           ********************************** ************************************************

SC-05  337:GGLRPGAPTLPPSTIVRPELSPGLKLTMQDLAVDKIVDTGIGFVSPDMANEFYEGLHSHLQNVGIDGVKVDVIHILEMLCEKYGGRV 423
SC-07  132:GGLRPGAPTLPPSTIVRPELSPGLKLTMQDLAVDKIIDTGIGFVSPDMANEFYEGLHSHLQNVGINGVKVDVIHILEMLCEKYGGRV 218
           *********************************** *************************** ********************

SC-05  424:DLAKAYFKALTSSVNKHFDGNGVIASMEHCNDFMFLGTEAISLGRVGDDFWCTDPSGDINGTYWLQGCHMVHCAYNSLWMGNFIQPDWD 512
SC-07  219:DLAKAYFKALTSSVNKHFDGNAVIASMEHCNDFMFLGTEAISLGRVGDDFWCTDPSGDINGTYWLQGCHMVHCAYNSLWMGNFIQPDWD 307
           ********************* *******************************************************************

SC-05  513:MFQSTHPCAEFHAASRAISGGPIYISDCVGQHDFDLLKRLVLPDGSILRCEHYALPTRDRLFEDPLHDGKTMLKIWNLNKYTGIIGAFNC 602
SC-07  308:MFQSTHPCAEFHAASRAISGGPIYISDCVGQHDFDLLRRLVLPDGSILRCEYYALPTRDRLFEDPLHDGKTMLKIWNLNKYTGIIGAFNC 397
           ************************************* ************* **************************************

          ->609
SC-05  603:QGGGWCRETRRNQCFSQCVNTLTATTNPKDVEWNSGNNPISVENVEEFALFLSQSKKLVLSGPNDDLEITLEPFKEELITVSPVVTI 689
SC-07  398:QGGGWCRETRRDQCFSQCVNTLTATTNPNDVEWNSGNNPISIENVEEFALFLSQSKKLVLSGQNDDLEITLEPFKFELITVSPVVTI 484
           *********** **************** *********** ******************** ************************

         695<-
SC-05  690:EGSSVQFAPIGLVNMLNTSGAIRSLVYHEESVEIGVRGAGEFRVYASRKPASCKIDGEVVEFGYEESMVMVQVPWSAPEG 769
SC-07  485:EGSSVQFAPIGLVNMLNTSGAIRSLVYHEESVEIGVRGAGEFRVYASKKPVSCKIDGEDVEFGYEESMVMVQVPWSAPEG 564
           *********************************************** ** ******* *********************

SC-05  770:LSSIKYEF                                                                                 777
SC-07  565:LSSIKYLF                                                                                 572
           ****** *
```

Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

JAMES U. BOWIE,* JOHN F. REIDHAAR-OLSON, WENDELL A. LIM, ROBERT T. SAUER

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

THE GENOME IS MANIFEST LARGELY IN THE SET OF PROteins that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is fundamental to an understanding of biological processes. Because an amino acid sequence contains all of the information necessary to determine the structure of a protein (*1*), it should be possible to predict structure from sequence, and subsequently to infer detailed aspects of function from the structure. However, both problems are extremely complex, and it seems unlikely that either will be solved in an exact manner in the near future. It may be possible to obtain approximate solutions by using experimental data to simplify the problem. In this article, we describe how an analysis of allowed amino acid substitutions in proteins can be used to reduce the complexity of sequences and reveal important aspects of structure and function.

Methods for Studying Tolerance to Sequence Variation

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. This method has been extremely powerful for proteins such as the globins or cytochromes, for which sequences from many different species are known (*2–7*). The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (*3, 8–11*). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions (*2–4, 11*). For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in *lac* repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent (*11*). At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a 10^4-fold reduction in activity (*12*). A similar loss of activity occurs in λ repressor when a DNA binding residue is changed from Asn to Asp (*13*). To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity (*10, 14–16*). Hence, many of the residues that are conserved in sets of related sequences play structural roles.

Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved (*6*). Similar results have been seen for a number of protein families (*2, 4, 5, 7, 17, 18*). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in λ repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the protein (*9*). These substitutions were identified by a functional

The authors are in the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

*Present address: Department of Chemistry and Biochemistry and the Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90024.

the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (*4*), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in λ repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of active sequences, but it is not possible to decide whether a side chain is functionally or structurally important just because it is invariant or conserved. To make this distinction requires an independent assay of protein folding. The ability of a mutant protein to maintain a stably folded structure can often be measured by biophysical techniques, by susceptibility to intracellular proteolysis (*26*), or by binding to antibodies specific for the native structure (*27, 28*). In the latter cases, it is possible to screen proteins in mutated clones for the ability to fold even if these proteins are inactive. Sets of sequences that allow formation of a stable structure can then be compared to the sets that allow both folding and function, with the active site or binding residues being those that are variable in the set of stable proteins but invariant in the set of functional proteins. The DNA-binding residues of Arc repressor were identified by this method (*8*). The receptor-binding residues of human growth hormone were also identified by comparing the stabilities and activities of a set of mutant sequences (*28*). However, in this case, the mutants were generated as hybrid sequences between growth hormone and related hormones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (*29, 30*). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have

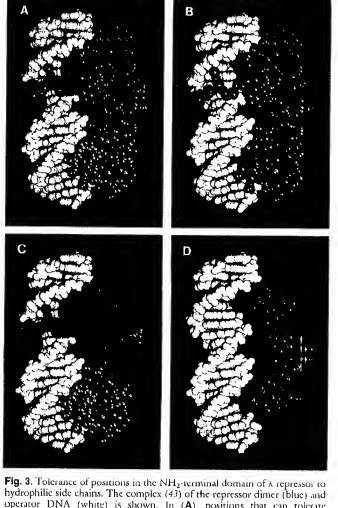


Fig. 3. Tolerance of positions in the NH_2-terminal domain of λ repressor to hydrophilic side chains. The complex (*43*) of the repressor dimer (blue) and operator DNA (white) is shown. In (**A**), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (**B**) without the remaining protein atoms. In (**C**), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (**D**) without the remaining protein atoms. About three-fourths of the 92 side chains in the NH_2-terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (*9, 14, 30, 37, 44*).

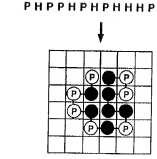


Fig. 5. A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (*40*), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, F. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965).
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. J. H. Miller *et al.*, *J. Mol. Biol.* **131**, 191 (1979).
12. S. Sprang *et al.*, *Science* **237**, 905 (1987); C. S. Craik, S. Rozeznak, C. Largman, W. J. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* **26**, 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and G. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (*5*) have argued that a protein core composed solely of hydrogen bonded residues would also be invisible on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989), W. S. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Breyer and R. T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330 (1989).
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown *et al.*, *J. Mol. Biol.* **42**, 65 (1969); J. Greer, *ibid.* **153**, 1027 (1981); J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia *et al.*, *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim, *J. Mol. Biol.* **88**, 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982); D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988).
36. G. Otting *et al.*, *EMBO J.* **7**, 4305 (1988).
37. L. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. Patel, *ibid.*, p. 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.*, in preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987); D. G. Covell and R. L. Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369 (1983).

46. We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).

ALGORITHMS FOR MULTIPLE SEQUENCE ALIGNMENTS

Guy Bottu,
BEN, The Belgian EMBnet node.

Introduction.

In a previous issue of embnet.news, we considered the alignment of pairs of sequences and the search for similar sequences in databanks. We now turn our attention to multiple sequence alignments.

If you have several similar nucleic acid or protein sequences it is often useful to align corresponding bases or amino acids in columns. For instance, you might wish to group bases or amino acids that occupy similar positions in the three-dimensional structure which exercise similar functions or that have evolved by substitution from the same base or amino acid in an ancestral sequence. In the latter case you might also like to construct a phylogenetic tree.

1. Global alignments.

The Needleman and Wunsch algorithm for finding the best global alignment of two sequences can readily be extended to multiple sequences. The problem is that the time the computer needs for such a job is roughly proportional to the product of the sequence lengths. So, if aligning two sequences of 300 positions takes 1 second, aligning 3 sequences takes 300 seconds and aligning 10 sequences would take 300^8 seconds, which is longer than the lifetime of the universe!

Since searching for a best global alignment using a rigorous algorithm is not realistic for more than three sequences, a number of strategies have been developed to carry out a multiple global alignment in a reasonable amount of time with a reasonable chance of finding the best alignment. The GCG program pileup first aligns all possible pairs of sequences according to Needleman and Wunsch (for n sequences, this makes n*(n-1)/2 alignments). Then it uses the pairwise similarity scores to construct a tree using the UPGMA method (see below). Finally, this tree serves as a guide for a progressive multiple alignment starting from the tips. Once two sequences have been aligned, their relative alignment is no longer changed. Clusters of previously aligned sequences are treated as a linearly weighted profile when they are subsequently aligned with another sequence or another cluster.

Other approaches include:

- The very popular CLUSTAL program differs only from pileup in that it performs the initial pairwise alignments using the fast algorithm of Wilbur and Lipman. CABIOS 8:189 (1992). References } ✓
  you can obtain versions of CLUSTAL for UNIX and for VAX
- Starting with a search for words of n bases or amino acids that are common between the sequences. An example is Martin Vingron's program MALI. CABIOS 5:115 (1989). References.
  MALI is not distributed freely but may be obtained from its author Martin Vingron (vingron@embl-heidelberg.de)
- PIMA uses pattern-matching, rather then profile matching, while making the progressive alignment. PNAS 87:118 (1990) References
  PIMA can be obtained for UNIX and for VAX
- Building a phylogenetic tree, using a more elaborate algorithm, as the sequences are progressively aligned. An example is Jotun Hein's program TreeAlign. Meth.Enzymol. 18:626(1990)
  TreeAlign can be obtained for UNIX and VMS from the same address as given for Clustal (see above)
- Making the best multiple alignment in a limited area of alignment space. This can only realistically be performed with eight to ten sequences.

2. Local alignments.

There are cases where sequences share a similar region but are otherwise completely different. Take, for example, the amino acids in the active site of an enzyme or transcription factor binding sites in a DNA sequence. To handle these cases local multiple alignment algorithms have been developed. Usually they only look for ungapped alignments thereby avoiding the problem of choosing the optimal gap penalty. Two such programs have been developed at the NCBI :
MACAW by Schuler, Altschul and Lipman first tries to find high scoring segment pairs (HSPs) for each possible pair of sequences using the BLAST algorithm (with the sensitivity set high). It then assembles overlapping HSPs into blocks. An interesting feature of MACAW is that it does not try to align all sequences, but can pick out only those that share similar regions. Proteins 9:180 (1991). References
There are versions of MACAW obtainable for the PC under Windows and for the Mac.
The MACAW distribution also contains Gibbs (see below) and a pattern searcher.

The Gibbs sampler algorithm involves iteratively making a profile with stretches of n bases or amino acids, selected from the sequences, and then searches this profile against one of the sequences. The result of the search is used to weight the selection of the stretches at the next run. A drawback is that the user must choose the width n and the number of elements in each sequence and thus must have a certain idea of the outcome, or run the program several times. An interesting feature is that the Gibbs sampler algorithm avoids the choice of an externally added scoring scheme since it derives the highest scoring profile, in a

self-consistent manner, from the data. Science 262:208 (1992). References .
Gibbs for is available for UNIX.

3. blast3.

It is also worth mentioning the program blast3. This searches a protein against a protein databank using the BLAST algorithm (with the sensitivity set high) and then makes threefold alignments between the query sequence and each possible pair of databank sequences that have been found. Only the statistically significant threefold alignments which are made from three nonsignificant pairwise alignments are retained. blast3 is useful in finding proteins that share a region of only weak similarity. Occasionally it can show that a query sequence makes the bridge between two databank sequences whose relationship had not yet been suspected.
You can look at the Manual.
It is possible to access a BLAST (including blast3) server at the NCBI, either through WWW or with a specific blast Internet client that you can install on your computer. More INFO is available.

4. phylogenetic trees.

Ideally a researcher would like to have a black box in which to throw sequences and get out a fully annotated phylogenetic tree. This is, however, not possible for two reasons. First, an algorithm that considers all possible multiple sequence alignments and then, for each alignment, all possible phylogenetic trees and picks out the best one, would take too much time. That is why most phylogenetic programs work on previously aligned sequences. Second, the result is always strongly influenced by the criteria that are used to define the best tree. Phylogenetic analysis will be the subject of a separate column in a later issue of embnet.news. However, a few remarks seem appropriate here. There are three main kinds of tree building methods: distance matrix, maximum likelihood and parsimony.

Distance matrix methods first estimate the pairwise distances between the sequences (which means that the information in the alignment of two sequences is reduced to one number) while the other methods construct many trees from all the information in the multiple alignment and decide which is best.

The simplest distance based method is UPGMA (unweighted pair-group method using arithmetic averages) which involves iteratively taking together the two sequences that have the shortest distance from each other, placing them at the end of branches on a node of the tree, and replacing their distances from the other sequences by an average value.

The guide tree used by pileup and CLUSTAL should never be used to infer phylogeny! It has been derived from the distances between pairwise aligned sequences and these distances are not necessarily the same as the distances between sequence pairs taken from the multiple sequence alignment.

Go to: previous article - next article - Table to contents